CISTER
Research Centre in
Real-Time & Embedded
Computing Systems

Conference Paper

Reliability Test based on a Binomial Experiment for Probabilistic Worst-Case Execution Times

Luìs Fernando Arcaro
Karila Palma Silva
Rômulo Silva de Oliveira
Luís Almeida

CISTER-TR-201103

Reliability Test based on a Binomial Experiment for Probabilistic Worst-Case Execution Times

Luìs Fernando Arcaro, Karila Palma Silva, Rômulo Silva de Oliveira, Luís Almeida

CISTER Research Centre

Polytechnic Institute of Porto (ISEP P.Porto)

Rua Dr. António Bernardino de Almeida, 431

4200-072 Porto

Portugal

Tel.: +351.22.8340509, Fax: +351.22.8321159

E-mail:

https://www.cister-labs.pt

Abstract

Measurement-Based Probabilistic Timing Analysis(MBPTA) produces Probabilistic Worst-Case Execution Times(pWCETs), i.e., WCET estimates associated with known lowexceedance probabilities. Despite applicability and goodness-offittests being used within MBPTA, any method based on thesampling of a population is subject to a degree of uncertainty.The acceptance of MBPTA in industrial engineering processes dependson obtaining enough evidence that the produced pWCETsare indeed reliable. In this paper we propose a statisticalhypothesis test to check the reliability of pWCET estimates, doneat a specified significance level. We assume as null hypothesis thatthe pWCET estimate is reliable, and as alternative hypothesis thatit is optimistic. Both Type I and Type II errors are considered.The reliability test is based on a binomial experiment and itis complementary to applicability and goodness-of-fit tests. Weevaluated the test using multiple synthetic and real-hardwareexecution time samples, and applied it on 20 pWCET estimatesgenerated for each of them. The combined use of the proposedreliability test with applicability and goodness-of-fit tests coulddetect most of the knowingly unreliable estimates on syntheticsamples. Similar behaviour was observed for real-hardwaresamples, evidencing the test 19s usefulness for selecting pWCETestimates with increased confidence.

Reliability Test based on a Binomial Experiment for Probabilistic Worst-Case Execution Times

Luís Fernando Arcaro, Karila Palma Silva
Instituto SENAI de Inovação em Sistemas Embarcados (ISI-SE)
Florianópolis, Brazil
{luis.arcaro, karila.silva}@sc.senai.br

Rômulo Silva de Oliveira
Universidade Federal de Santa Catarina (UFSC)
Florianópolis, Brazil
romulo.deoliveira@ufsc.br

Luís Almeida
CISTER / Fac. Engenharia Universidade do Porto (UP)
Porto, Portugal
lda@fe.up.pt

***Abstract*—Measurement-Based Probabilistic Timing Analysis (MBPTA) produces Probabilistic Worst-Case Execution Times (pWCETs), i.e., WCET estimates associated with known low exceedance probabilities. Despite applicability and goodness-of-fit tests being used within MBPTA, any method based on the sampling of a population is subject to a degree of uncertainty. The acceptance of MBPTA in industrial engineering processes depends on obtaining enough evidence that the produced pWCETs are indeed reliable. In this paper we propose a statistical hypothesis test to check the reliability of pWCET estimates, done at a specified significance level. We assume as null hypothesis that the pWCET estimate is reliable, and as alternative hypothesis that it is optimistic. Both Type I and Type II errors are considered. The reliability test is based on a binomial experiment and it is complementary to applicability and goodness-of-fit tests. We evaluated the test using multiple synthetic and real-hardware execution time samples, and applied it on 20 pWCET estimates generated for each of them. The combined use of the proposed reliability test with applicability and goodness-of-fit tests could detect most of the knowingly unreliable estimates on synthetic samples. Similar behaviour was observed for real-hardware samples, evidencing the test's usefulness for selecting pWCET estimates with increased confidence.**

***Index Terms*—Real-time systems, timing analysis, worst-case execution time, embedded software.**

I. Introduction

Real-Time Systems (RTS) are subject to timing constraints, and vary greatly in scale, complexity and criticality. Timing constraints in these systems are represented as deadlines for completing software tasks, which must be met even in the worst-case scenario. A fundamental design step is to calculate upper-bound estimates of the longest possible time taken by the target hardware to execute each task, i.e., to determine the tasks' Worst-Case Execution Time (WCET).

Multiple methods exist for the determination of WCET upper bounds [1]. Static analysis generates reliable results through a detailed inspection of the task's code and of the hardware architecture, but the low time composability of current programs executing on complex modern processor architectures may exclude its use in practice.

This work has been partially supported by CAPES, The Brazilian Agency for Higher Education, project PrInt CAPES-UFSC "Automation 4.0", and by FCT/MCTES (PIDDAC), The Portuguese Foundation for Science and Technology, with CISTER Research Unit Base Funding (UIDB/04234/2020).

Measurement-based methods analyse tasks' execution times effectively yielded at run time, hence reducing analysis efforts and being potentially applicable to complex architectures. But these methods require setting safety margins to account for possibly unobserved timing events.

Measurement-Based Probabilistic Timing Analysis (MBPTA) determines WCET upper bounds based on the statistical analysis of execution time measurements [2]. It applies Extreme Value Theory (EVT), a statistical framework designed for estimating the probability of unusual events by modelling extreme deviations from the typical behaviour of the analysed phenomena [3]. MBPTA produces Probabilistic Worst-Case Execution Times (pWCETs), i.e., WCET estimates that have a probability of being exceeded and that probability can, in principle, be set to arbitrarily low values.

For instance, the electric power industry and the automotive industry use mission-critical equipment whose correct timing behaviour is usually assessed only by conventional testing. Due to the complexity of the hardware and software involved, it may not be feasible to use classic static timing analysis. But the use of MBPTA in such scenarios might be feasible if sufficient evidence of its correctness is provided.

The application of EVT in the context of MBPTA requires fitting an adequate extreme value distribution to the maximum execution times measured while the task is executed on the target hardware platform. This distribution is then used for determining a $pWCET(\epsilon)$, where ϵ represents an associated target exceedance probability. The proper application of EVT requires the measurements to present certain statistical properties. Also, the fitted model distribution must pass goodness-of-fit tests regarding its adjustment to the actual measurements [3].

Despite the applicability and goodness-of-fit tests, any method based on the sampling of a population is subject to a degree of uncertainty. The acceptance of MBPTA in industrial engineering processes depends on obtaining enough evidence that the resulting pWCET is indeed reliable.

In this paper we define a statistical hypothesis test to check the reliability of the resulting $pWCET(\epsilon)$. We assume that proper applicability and goodness-of-fit tests were applied and passed, so the null hypothesis is that $pWCET(\epsilon)$ is reliable, i.e., the actual exceedance probability is lower than or equal to ϵ. The alternative hypothesis is that $pWCET(\epsilon)$ is optimistic,

i.e., the actual exceedance probability is greater than ϵ.

The reliability test is based on modelling the exceedance of a concrete $pWCET(\epsilon)$ estimate as a binomial experiment with a validation sample. Our test is complementary to applicability and goodness-of-fit tests. The test is done at a specified significance level, which defines the probability of Type I error, i.e., the probability of rejecting as unreliable a $pWCET(\epsilon)$ that is actually reliable. The power of the test and the probability of Type II error are also described.

The usefulness of the reliability test proposed in this paper is illustrated with both synthetic and real-hardware examples. As synthetic data we used 5 samples drawn from EVT-compliant distributions [4], and generated 20 $pWCET(\epsilon)$ estimates for each of them. The combined use of the reliability test with applicability and goodness-of-fit tests could eliminate most estimates that were known to be unreliable. When applied to real-hardware measurements, the same evaluation procedure led to similar behaviour regarding the pWCET estimates produced using MBPTA, evidencing that the proposed test is useful for selecting $pWCET(\epsilon)$ estimates with increased confidence for use in engineering projects.

As an additional benefit, the proposed reliability test can also be incorporated into the product for continuous execution during the product's lifetime within an early fault detection system. To the best of the authors' knowledge, this is the first time that a statistical hypothesis test is used to assess the reliability of pWCET estimates.

The remainder of the paper is organized as follows. Section II presents the background of MBPTA and EVT and the corresponding related work. Section III introduces statistical hypothesis testing and its terminology and methods. Section IV presents our core contribution, i.e., the use of a binomial experiment to support a reliability test of pWCET estimates. Section V introduces a second contribution, which is the use of the statistical power of the test to characterize the probability of the proposed test rejecting unreliable pWCET estimates. Sections VI and VII validate the proposed approach on synthetic tasks and on concrete functions executing on actual hardware. Finally, Section VIII presents the concluding remarks.

II. Background

MBPTA is a timing analysis technique that aims at determining probabilistically reliable WCET bounds for real-time tasks, through the statistical analysis of execution time measurements. MBPTA applies EVT to measurements of the execution time of the task running on its target environment. An asymptotic distribution of extreme values is then adjusted to the measurements and used to produce the bounds (see the surveys [5] and [6]). WCET bounds produced by MBPTA – known as $pWCET(\epsilon)$ – are composed of both an upper-bounding value and a probability ϵ of that value being exceeded in any individual execution of the task [5].

To apply EVT, observed execution times must be amenable to representation as independent and identically distributed (i.i.d.) random variables. Recent work [7] supports EVT applicability as long as the measurements come from a stationary and identical distribution and there is extremal independence.

The assumptions for EVT applicability can be tested using several techniques ([8], [9], [10] and [11]), such as:

- **WW:** Wald-Wolfowitz test of independence.
- **TP:** Turning Point test of randomness.
- **LB2:** Ljung-Box test of absence of correlation between observed values with a lag of 2.
- **LB10:** Ljung-Box test with a lag of 10.
- **KS:** Kolmogorov-Smirnov test of identical distribution.
- **AD1:** Anderson-Darling test of identical distribution, version 1 (adjusts for possibly different sample sizes).
- **AD2:** Anderson-Darling test of identical distribution, version 2 (focuses on tail differences).

These tests are expected to produce so-called *p-values* distributed in $[0, 1]$, with failure indicated by a tendency to low values (e.g., < 0.05). Results are presented as box and whisker plots highlighting quantiles 0%, 5%, 50%, 95% and 100%.

The two main EVT approaches to sample extremes are Block Maxima (BM) and Peaks over Threshold (POT), neither of them dominating the other. In our work we use Block Maxima. The application of EVT in the context of MBPTA using the BM approach requires the following steps:

1) Measuring the execution time of several runs of the task.
2) Providing evidence that EVT can be applied.
3) Dividing the sample into identically sized blocks and retaining only each block's largest value.
4) Fitting a Generalized Extreme Value GEV(μ, σ, ξ) distribution to the measurements by estimating its parameters location (μ), scale (σ) and shape (ξ).
5) Testing goodness-of-fit between the measurements and the fitted distribution, using e.g. plots or statistical tests.
6) Using the fitted model to obtain a $pWCET(\epsilon)$ value with the desired exceedance probability ϵ (or vice-versa).

The uncertainty of any method based on sampling is usually dealt with in statistical approaches by using confidence intervals and confidence levels. Regarding MBPTA, the testing of applicability conditions may leave lingering doubt [4]. Moreover, many fitting methods provide confidence intervals for the parameters of the adjusted model distribution with a confidence level that can be arbitrarily set, but which is usually fixed to 95%. The actual value of those parameters can occasionally fall outside the confidence interval. Also, several statistical tests are affected by the sample size [12]. These factors may lead such methods to yield unreliable, i.e. optimistic, pWCET estimates. In an RTS, a $pWCET(\epsilon)$ estimate is reliable when its estimated exceedance probability ϵ is equal to or greater than the actual (unknown) probability of the analysed task yielding longer execution times.

When EVT is applied to natural phenomena, such as river levels and rain volumes [13], the available data is limited and the reliability of the exceedance probabilities of worst-case estimates can be evaluated only with the passage of time. However, it is possible to obtain much larger samples with computing systems than what is achievable with natural

phenomena. Part of these samples can hence be used to assess the reliability of a concrete $pWCET(\epsilon)$. Large samples have been used before to evaluate the variability of pWCET estimates [2], to assess pessimism [14] and to evaluate the impact of block sizes and of POT thresholds [10].

The work in [15] evaluated the reliability of pWCET estimates comparing the estimates with the known real WCET of synthetic tasks executed on a controlled abstract platform. The subsequent works in [16]–[18] evaluated reliability by comparing $pWCET(\epsilon)$ with the HWM (High-Water Mark) of a validation sample with n measurements, i.e., the highest execution time value observed in a large sample. Although a valid method, the reliability test described in this paper is superior in the sense that, depending on the values of ϵ and n, the observance of a few exceedances may not be sufficient to deem a $pWCET(\epsilon)$ unreliable. Taking that into account, the work in [19] defines an Exceedance Density Metric by comparing the frequencies of exceedances to their expected average value. In this paper we use a similar principle but propose, instead, a proper statistical hypothesis test, explicitly considering both Type I and Type II errors (Section III).

The work described in [4] analysed and compared statistical tests for verifying the applicability requirements of EVT. In [20] the authors present the statistical power estimation of several goodness-of-fit tests for EVT distributions. Differently from those two works, in this paper we test the final product of MBPTA, i.e., pWCET estimates, and not the applicability requirements or the goodness-of-fit of the model distribution. The reliability test presented in this paper is complementary to already established MBPTA applicability tests, e.g., stationarity, independence, and goodness-of-fit tests.

Recent works [21] aim at leveraging the application of MBPTA in the context of critical systems, a scenario in which reliability tests such as the one proposed in this work may prove particularly useful.

III. Statistical Hypothesis Tests

A statistical hypothesis is a claim (assumption) about an unknown parameter of a population [22]. H_0 denotes the null hypothesis, which is assumed to be true before the test. The alternative hypothesis is denoted by H_1 and it contradicts the null hypothesis. The test is a decision rule used to decide whether sufficient evidence exists on a sample to reject the null hypothesis H_0 in the population.

In statistical hypothesis testing we select a random sample of the population, compute a sample statistic related to the population parameter of interest, and then judge its consistency with H_0. The statistic computed for the sample is called the test statistic. We compare the test statistic of the sample with what would be expected in case H_0 was actually true. H_0 is rejected when it is very unlikely to observe such sample statistic value if H_0 is actually true.

Hypotheses are always statements about the population under study, not statements about the sample. The truth or falsity of a particular hypothesis can never be known with certainty unless we can examine the entire population. For this reason, hypothesis testing provides evidence, not certainty.

Two errors are inherent to statistical hypothesis tests:

- Reject H_0 when it is actually true (Type I error).
- Do not reject H_0 when it is actually false (Type II error).

We use α to represent the probability of Type I error in a statistical hypothesis test, i.e., the probability of rejecting H_0 when it is actually true. α is also called *significance level*. β represents the probability of Type II error in a hypothesis test, i.e., the probability of not rejecting H_0 when it is actually false. The *power of the test* is defined as $(1-\beta)$. Table I summarizes the error types in statistical hypothesis tests.

TABLE I
Error types in statistical hypothesis tests

Decision	H_0 **True,** H_1 **False**	H_0 **False,** H_1 **True**
Do Not Reject H_0	Correct Decision $(1-\alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	Correct Decision $(1-\beta)$

The choice of α is always a decision of the analyst and it reflects in part the confidence the analyst has in H_0 before the test. Let's suppose the analyst has great confidence that H_0 is true, due to previous experience and tests. The analyst will reverse that belief and reject H_0 if the hypothesis test provides strong evidence that H_0 is actually false, only. In this case, it is appropriate to use a small α, such as 1%, 0.1% or even smaller values. The analyst will reject H_0 only if the test outcome has a probability of, say, 0.1% or less of being observed, was H_0 true. That will be the probability of rejecting a true H_0, i.e., the probability of Type I error. At the same time, being α set to such low values makes it easy failing to reject a false H_0, since the rejection requires very strong evidence.

Now let us suppose the analyst believes H_0 is true, but with a limited level of confidence. In this case, it is appropriate to use the typical value of 5% for α. The analyst will reject H_0 when the test outcome has a probability of 5% or less of being observed, was H_0 true. Now it becomes more likely to reject a false H_0, because the rejection criterion is less demanding.

In the context of RTS and MBPTA, when H_0 is "$pWCET(\epsilon)$ is reliable", to reject a true H_0 is undesirable but safe. On the other hand, not rejecting a false H_0 is unsafe. In this case an α of 5% is more appropriate than, say, 0.1% or even 1%.

A statistical hypothesis test is called unilateral when the rejection comes from the sample statistic being too big, or being too small, but only one of them. Such test is also referred to as one-tailed or one-sided. It is called a bilateral test when the rejection comes from the sample statistic being either too big or too small, any one will do. Such test is also referred to as two-tailed or two-sided.

Let $E(m)$ denote the random variable that represents the test statistic, where m is some parameter used in the test. e is a random variate, i.e., a particular outcome, of $E(m)$. There are two approaches to hypothesis testing.

The *classic approach* includes the following steps:

1) Define the significance level α.
2) Compute the critical value c, so there is a probability α of obtaining a sample statistic at least as extreme as the critical value c, assuming that the null hypothesis is true (right-tailed test): $\alpha = P(E(m) \geq c \mid H_0)$.
3) Compute the statistic value e for the sample.
4) Compare e with the critical value c in order to decide whether H_0 should or should not be rejected.
5) We reject H_0 when $e \geq c$.
6) We do not reject H_0 when $e < c$.
7) The critical region corresponds to the set of all values of the sample statistic $E(m)$ that will lead to rejection of H_0 at significance level α.

The *p-value approach* includes the following steps:

1) Define the significance level α.
2) Compute the statistic value e for the sample.
3) Compute the *p-value*, which consists of the probability of obtaining a sample statistic at least as extreme as e, assuming the null hypothesis is true (right-tailed test): *p-value* $= P(E(m) \geq e \mid H_0)$.
4) Compare the *p-value* with α in order to decide whether H_0 should or should not be rejected.
5) We reject H_0 when *p-value* $\leq \alpha$.
6) We do not reject H_0 when *p-value* $> \alpha$.
7) We reject H_0 when the obtained *p-value* is too small in relation to α, i.e., when it is unlikely we would observe that sample statistic value if H_0 was true.

Both hypothesis testing approaches are equivalent, since they lead to the same decision under the same conditions. Throughout this work we use the *p-value approach*, because it seems easier for the reader to understand its semantics.

A. Binomial Experiment

A statistical experiment is called a binomial experiment if (i) it consists of n independent trials, (ii) each trial can lead only to two possible outcomes: a success or a failure, and (iii) the probabilities of a success and of a failure are given respectively by p and q, such that $p + q = 1$ [22].

In a binomial experiment, the probability of exactly k successes being observed in n trials, assuming p and q are in fact respected, can be calculated using the probability mass function of the binomial distribution given by:

$$pmf(k, n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

Similarly, the probability of at most k successes being observed in n trials can be calculated using its cumulative distribution function given by $cdf(k, n, p) = \sum_{i=0}^{k} pmf(i, n, p)$. Since the Binomial Distribution is discrete, we must observe that $cdf(k, n, p) = cdf(k-1, n, p) + pmf(k, n, p)$.

The binomial experiment theory can be used to model and reason about numerous phenomena, from which we use as example the 100-year flood [23]. The 100-year flood is a flood event whose probability of being equalled or exceeded in any year is 0.01 (i.e. its exceedance probability is 1%). In this regard, a very common misunderstanding that can be clarified using the binomial experiment theory is that a 100-year flood is likely to occur only once in every period of 100 years. Despite it is correct to affirm that 100-year floods occur *in average* once in every period of 100 years, by modelling the phenomenon as a binomial experiment it can be shown that the probability of at least one 100-year flood being observed within a randomly chosen period of 100 years is actually $1 - cdf(0, 100, 0.01) \approx 63.40\%$. Moreover, the probability of at least two of such events occurring within a 100-year window is of approximately 26.42%. If we reject a 100-year flood estimate just because it was exceeded twice in a period of 100 years, the chance of being wrong (Type I error, i.e. rejecting a true null hypothesis) would be of 26.42%.

IV. Reliability Test Based on a Binomial Experiment

Assume $pWCET^*(\epsilon)$ denotes the actual pWCET of a task with an exceedance probability of ϵ, which is a value usually unknown. Let $pWCET^j(\epsilon)$ denote the *j-th* estimate of the task's pWCET with a target exceedance probability of ϵ, which is obtained through the application of MBPTA. The superscript j can be omitted when there is a single estimate.

Given the nature of the fitting process, it is unlikely that it will produce an estimate $pWCET(\epsilon)$ that is exactly the value of $pWCET^*(\epsilon)$. Without loss of generality, let's assume that the actual exceedance probability of $pWCET(\epsilon)$ is ω, i.e., $pWCET(\epsilon) = pWCET^*(\omega)$. Observe that, by the definition of pWCET, $\omega > \epsilon \Rightarrow pWCET^*(\omega) < pWCET^*(\epsilon)$ and that $\omega < \epsilon \Rightarrow pWCET^*(\omega) > pWCET^*(\epsilon)$.

We say $pWCET(\epsilon)$ is optimistic (unreliable) when the estimate is smaller than the actual value, i.e., $pWCET(\epsilon) < pWCET^*(\epsilon)$. In this case we have $pWCET(\epsilon) = pWCET^*(\omega) < pWCET^*(\epsilon)$ and $\omega > \epsilon$.

We say $pWCET(\epsilon)$ is reliable when the estimate is greater than or equal to the actual value, i.e., $pWCET(\epsilon) \geq pWCET^*(\epsilon)$. In this case we have $pWCET(\epsilon) = pWCET^*(\omega) \geq pWCET^*(\epsilon)$ and $\omega \leq \epsilon$.

For example, suppose the execution times of a task are such that there is a probability of 10^{-9} of it being greater than 1000 clock cycles, and a probability of 10^{-8} of it being greater than 900 clock cycles, i.e., $pWCET^*(10^{-9}) = 1000$ and $pWCET^*(10^{-8}) = 900$. These values are usually unknown. Assume we are interested in the target exceedance probability of $\epsilon = 10^{-8}$, so we applied MBPTA to this task and obtained an estimate $pWCET(10^{-8}) = 1000$. We say this estimate is reliable because $pWCET(10^{-8}) \geq pWCET^*(10^{-8})$. We also have $pWCET(10^{-8}) = pWCET^*(10^{-9}) \geq pWCET^*(10^{-8})$ and $\omega = 10^{-9} \leq \epsilon = 10^{-8}$.

In this paper we define a statistical test for the reliability of $pWCET(\epsilon)$, where ϵ represents the targeted exceedance probability. A pWCET estimate is considered reliable when its targeted exceedance probability ϵ is higher than or equal to the actual (unknown) probability ω of the task under analysis yielding longer execution times. We assume the estimate $pWCET(\epsilon)$ was obtained through a careful application of EVT and that proper applicability and goodness-of-fit tests

were applied and passed. So, the null hypothesis H_0 is that $pWCET(\epsilon)$ is reliable:

H_0: $pWCET(\epsilon) = pWCET^*(\omega), \omega \leq \epsilon$

H_1: $pWCET(\epsilon) = pWCET^*(\omega), \omega > \epsilon$

The alternative hypothesis H_1 is that $pWCET(\epsilon)$ is optimistic and the actual exceedance probability is greater than ϵ. The rejection of $pWCET(\epsilon)$ comes only from it being too small, so this is a *unilateral test*.

The pWCET exceedance problem can be modelled as a binomial experiment, in which each trial is associated with an individual execution of the task and is considered successful (regarding the binomial experiment) if and only if $pWCET(\epsilon)$ *is* exceeded in that particular execution. The experiment's success probability p hence equals the intended exceedance probability ϵ of the pWCET estimate $pWCET(\epsilon)$.

By using a validation sample V of n measurements and comparing each one with $pWCET(\epsilon)$, we have a set of n independent trials of the experiment. Let e denote the number of measurements that exceed $pWCET(\epsilon)$ in V. e is the test statistic, i.e., the outcome of $E(pWCET(\epsilon))$ for the test.

The proposed test's *p-value*, explained in Section III, comes from the right tail of the cumulative distribution function of a Binomial Distribution with n trials and success probability ϵ, considering e and greater values. It hence consists of the probability of observing at least e exceedances in n executions of the analysed task assuming that $pWCET(\epsilon) = pWCET^*(\epsilon)$, i.e. that $pWCET(\epsilon)$ is reliable. Formally:

$$\textit{p-value} = P(E(pWCET(\epsilon)) \geq e \mid H_0)$$

The probability of exactly e pWCET exceedances being observed purely by chance in a sample of n execution times, assuming that the exceedance probability $p = \epsilon$ is indeed respected, can be calculated using $pmf(e, n, p)$. Consequently, the probability of at least e exceedances being observed purely by chance under the same assumptions, i.e. the *p-value* of the reliability test, is given numerically by:

$$\textit{p-value} = 1 - \textit{cdf}(e - 1, n, p)$$

A. Definition of H_0

The reliability test described in this paper associates H_0 with "$pWCET(\epsilon)$ is reliable" and searches for evidence that it is actually unreliable. One could ask why not defining H_0 the other way around, that is "$pWCET(\epsilon)$ is unreliable," and then to search for evidence that it is actually reliable. There are two reasons to define H_0 as "$pWCET(\epsilon)$ is reliable."

First, the reliability test is applied to a pWCET estimate obtained through EVT after appropriate testing of the requirements, i.e., stationarity and identical distribution, extremal independence, and goodness-of-fit. At this point, previous experience indicates the $pWCET(\epsilon)$ estimate is indeed reliable, so this should be the null hypothesis.

Secondly, typical exceedance probabilities targeted in real-time systems are very small, such as 10^{-10}. If we define H_0 as "$pWCET(\epsilon)$ is unreliable," in order to reject it (and to produce evidence that it is actually reliable) it would be necessary to observe extremely large samples with a number of exceedances much smaller than the predicted.

For instance, when testing an estimate $pWCET(10^{-10})$ with a validation sample of 10^8 measurements, the probability of observing zero exceedances is 99%. In this case, in order to reject "$pWCET(\epsilon)$ is unreliable," it is necessary to observe much less than zero exceedances, which is impossible. For the test to be useful we would need a much greater validation sample, but notice that 10^8 measurements already represents a huge measurement effort. On the other hand, since the target exceedance probability is very small, observing a few exceedances in a relatively small validation sample is enough to reject the hypothesis "$pWCET(\epsilon)$ is reliable."

The reliability test described in this paper assumes that the pWCET estimate was obtained through careful application of EVT, passing all requirements' tests, so there is reason to believe it is reliable. The reliability test works as a safeguard to assess and reinforce the safety of the pWCET estimate, which is a crucial factor to the acceptance of MBPTA in the development process of critical RTS. In this sense, the reliability test described in this work is complementary to the testing of EVT applicability requirements.

B. Numerical Examples

Assume we applied MBPTA to task τ using two different samples and obtained two estimates of $pWCET^*(10^{-10})$, denoted by $pWCET^1(10^{-10})$ and $pWCET^2(10^{-10})$. We are going to apply the reliability test with a significance level of 5%, i.e., $\alpha = 0.05$. We compare both estimates to each execution time of a validation sample of size $n = 10^8$.

Estimate $pWCET^1(10^{-10})$ was exceeded $e_1 = 2$ times. The *p-value* for $pWCET^1(10^{-10})$ is:

$$1 - \textit{cdf}(e_1 - 1, 10^8, 10^{-10}) = 0.00005$$

This means that observing at least 2 exceedances in a sample of size $n = 10^8$, when the average rate of one exceedance every 10^{10} is indeed being respected, is associated with a probability of 0.005%. Thus, we must reject $pWCET^1(10^{-10})$ with a significance level of 5%. We would actually reject it even at a much lower significance level, such as 0.1%.

Estimate $pWCET^2(10^{-10})$ was exceeded $e_2 = 0$ times. The *p-value* for $pWCET^2(10^{-10})$ is:

$$1 - \textit{cdf}(e_2 - 1, 10^8, 10^{-10}) = 1$$

We cannot reject $pWCET^2(10^{-10})$ at a significance level of 5%, neither at any useful significance level.

Table II considers the case of $\epsilon = 10^{-10}$ and shows the probability of observing at least a certain number of exceedances (from 1 to 5, in each line) in a validation sample of a certain size (from 10^7 to 10^{11}). For example, in a sample of 10^{10} measurements there is a probability of 8% of observing at least 3 exceedances. In other words, if we reject $pWCET(10^{-10})$ because there are three exceedances in a sample of 10^{10} measurements, the probability of rejecting a

reliable $pWCET(10^{-10})$ is 8% (Type I error). One can easily compute tables like this for any value of ϵ.

TABLE II
PROBABILITY OF OBSERVING EXCEEDANCES FOR $pWCET(10^{-10})$

Number of exceedances	Probability by Validation sample size				
	10^7	10^8	10^9	10^{10}	10^{11}
≥ 1	0.001	0.010	0.095	0.632	> 0.999
≥ 2	$< 10^{-6}$	$< 10^{-4}$	0.005	0.264	> 0.999
≥ 3	$< 10^{-9}$	$< 10^{-6}$	< 0.001	0.080	0.997
≥ 4	$< 10^{-13}$	$< 10^{-9}$	$< 10^{-5}$	0.019	0.990
≥ 5	$< 10^{-17}$	$< 10^{-12}$	$< 10^{-7}$	0.004	0.971

As already referred, assuming EVT was carefully applied, and the fitting process passed all applicability and goodness-of-fit tests, there is reason to believe the estimated pWCET is reliable. Notwithstanding, given the importance of the pWCET for the safety of the designed system, we apply the reliability test with a significance level of 5%, expressing a limited level of confidence on its factual reliability.

Table III shows the critical value of the test statistic, i.e., the minimum number of exceedances one must observe in order to reject the pWCET estimate with a significance level of 5%. Each line shows a different value for ϵ and each column a different value for n (validation sample size). For example, it shows that $pWCET(10^{-10})$ can only be rejected with a significance level of 5% when we observe at least 4 exceedances in a validation sample of size 10^{10}.

TABLE III
CRITICAL VALUES AT 5% SIGNIFICANCE LEVEL

pWCET(ϵ)	Critical value by Validation sample size				
	10^6	10^7	10^8	10^9	10^{10}
$pWCET(10^{-7})$	2	4	16	118	1053
$pWCET(10^{-8})$	1	2	4	16	118
$pWCET(10^{-9})$	1	1	2	4	16
$pWCET(10^{-10})$	1	1	1	2	4
$pWCET(10^{-11})$	1	1	1	1	2
$pWCET(10^{-12})$	1	1	1	1	1

V. STATISTICAL POWER OF THE TEST

The statistical power of a hypothesis test is the probability of the test rejecting the null hypothesis H_0 when it is false and the alternative hypothesis H_1 is true.

There are several factors that improve the power of the test:

- The amount by which the null hypothesis is false.
- The larger the sample size, the larger the power.
- The larger the significance level, the larger the power.
- A one-tailed hypothesis puts a larger rejection region on the side of the true state of the world, increasing power.

Let ω be the actual value of the population parameter, which is usually unknown. In our case, ω represents the actual exceedance probability of the estimated $pWCET(\epsilon)$: $pWCET(\epsilon) = pWCET^*(\omega) < pWCET^*(\epsilon)$, with $\omega > \epsilon$. Recall that $\omega > \epsilon \Rightarrow pWCET^*(\omega) < pWCET^*(\epsilon)$.

Since ω is usually unknown, we describe the power of the test as $B(\omega)$, i.e., the probability of rejecting H_0 when H_1 is true and the true exceedance probability of our estimate is ω. We can plot $B(\omega)$ against ω in order to visualize the probability of rejecting a false H_0 depending on how "far" ω is from ϵ. The Type II error is given by $\beta(\omega) = 1 - B(\omega)$.

In order to compute $B(\omega)$ we first compute the critical value c, so that the probability of obtaining a sample statistic at least as extreme as c is less than or equal to α, assuming H_0 is true. This is the usual rejection criterion:

$$P(E(pWCET^*(\epsilon)) \geq c \mid H_0) \leq \alpha$$

We can compute $B(\omega)$ for various values of ω using:

$$B(\omega) = P(E(pWCET(\epsilon)) \geq c \mid pWCET(\epsilon) = pWCET^*(\omega)) \quad (1)$$

which is the probability of rejecting H_0 given that $pWCET(\epsilon) = pWCET^*(\omega) < pWCET^*(\epsilon)$, i.e., H_1 is true.

For example, assume we applied MBPTA to task τ and obtained an estimate $pWCET(\epsilon)$, $\epsilon = 10^{-10}$, which was $pWCET(10^{-10}) = 52000$. We are going to apply the reliability test with a significance level of 5%, i.e., $\alpha = 0.05$. We compared the estimate to a validation sample of size $n = 10^8$, where the maximum measurement was 50000. $pWCET(10^{-10})$ was exceeded $e = 0$ times.

The critical value is $c = 1$, since for $n = 10^8$ we have that $P(E(pWCET^*(10^{-10})) \geq 1) \leq 0.05$ from Table III. So, we cannot reject H_0 for $pWCET(10^{-10}) = 52000$.

Since $\epsilon = 10^{-10}$, let's analyse the power of the test for $\omega = 10^{-9}$. Assuming $pWCET(10^{-10}) = pWCET^*(10^{-9})$ and using Equation 1 we have that:

$$\begin{aligned} B(10^{-9}) &= P(E(pWCET^*(10^{-9})) \geq 1) \\ &= 1 - cdf(0, 10^8, 10^{-9}) \\ &= 0.09516 \end{aligned}$$

In this scenario, the probability of rejecting a false H_0 is $\approx 9.5\%$. We can repeat the process for other probabilities $\omega > 10^{-10}$, e.g., $B(10^{-8}) = 0.63212$ and $B(10^{-7}) = 0.99995$. The probability of rejecting a false H_0 is 99.995% if in this H_0 we have $pWCET(10^{-10}) = pWCET^*(10^{-7})$. Fig. 1 shows $B(\omega)$ against ω for $n = 10^8$ and $\alpha = 0.05$.

One can ask the question the other way around and determine which value of n will result in a probability of, say, 99% for rejecting a false H_0 if in that H_0 we have $pWCET(10^{-10}) = pWCET^*(10^{-9})$, i.e. if the $pWCET(10^{-10})$ estimate is in fact exceeded with a probability of 10^{-9}. The value of n affects both the critical value c and the power of the test, but the resulting problem can be solved iteratively. Assuming $\alpha = 0.05$, $\epsilon = 10^{-10}$ and $\omega = 10^{-9}$, we need $n = 1.005 \times 10^{10}$ to obtain $B(\omega) = 0.99$. With a sample size of $n = 3 \times 10^9$ we have $B(\omega) = 0.80$ in the same conditions.

VI. EXAMPLE WITH SYNTHETIC DATA

In order to illustrate the application of the proposed reliability test, we consider synthetic time samples from five hypothetical tasks. The execution time measurements are actually random

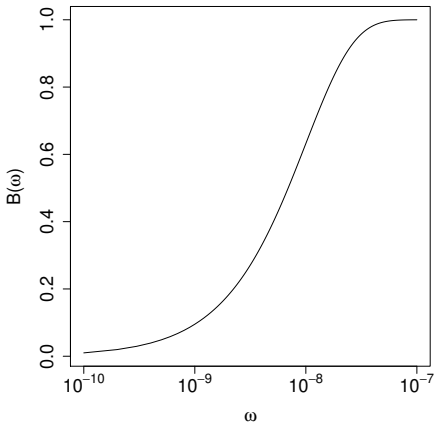


Fig. 1. Power of the Test $B(\omega)$ for $n = 10^8$ and $\alpha = 0.05$

numbers from EVT-compliant distributions [4], i.e., they are known to be independent and identically distributed. We use the following synthetic tasks:

- τ_p: Poisson distribution with expected value 10000;
- τ_{n1}: Normal distribution with $\mu = 10000$ and $\sigma = 200$;
- τ_{n2}: Normal distribution with $\mu = 10000$ and $\sigma = 1000$;
- τ_{g1}: Gamma distribution with shape 10000 and scale 1;
- τ_{g2}: Gamma distribution with shape 10000 and scale 3.

We used the Block Maxima approach with a GEV distribution fitted, using the L-moments method [24], to the measurements' maxima selected from blocks of size 100. The use of blocks of size 100 is common in pWCET literature, it has been shown adequate in previous work [17], and the maxima obtained from our samples passed MBPTA tests.

We collected a modelling sample of 200,000 measurements for each task. Then we made 20 fittings for each task, using for each fitting a different subsample of 10,000 measurements. Although there is an ongoing debate about which method to use when applying EVT in the context of RTS, the method and sample size used in this paper are similar to other works in the literature. Note that the purpose of the evaluations presented in this section is illustrating the application and highlighting the effectiveness and usefulness of the proposed reliability test. No focus is hence given to more general aspects of MBPTA application, such as the choice of sample size or of the probabilistic models to be used.

As an example, throughout this section we employ pWCET estimates with exceedance probability $\epsilon = 10^{-10}$, a validation sample with $n = 10^8$ measurements, and $\alpha = 0.05$. Since the conclusions are similar for all tasks, for the sake of simplicity we present detailed results only for tasks τ_{g2}, τ_p and τ_{n2}.

Table IV shows the dispersion of $pWCET^j(10^{-10})$, obtained from 20 different samples of τ_{g2}. Values of pWCET are presented in ascending order for ease of reading, but they could be obtained in any order. For each fitting of τ_{g2} it shows the $pWCET(10^{-10})$, the number e of measurements in the validation sample that exceed each pWCET estimate, and the *p-values* obtained from the two versions (AD1 and AD2) of the Anderson-Darling test implemented in the kSamples package of the R statistical software, which we use as goodness-of-fit (GOF) test [11]. Using a significance level of 5%, all 20 fittings would pass the GOF test. Throughout the paper we mark in bold when the *p-value* from the GOF test would result in the fitting being rejected at a significance level of 5%.

We applied the reliability test described in Section IV to the 20 fittings of τ_{g2}. Since all fittings passed the GOF test, for each one we have as null hypothesis that it is reliable. Table IV also shows, for each fitting of τ_{g2}, the *p-value* obtained when applying the reliability test. For instance, using a significance level of 5%, the null hypothesis would be rejected for fittings 1 to 9. Throughout the paper we mark in bold when the *p-value* from the reliability test results in the null hypothesis being rejected at a significance level of 5%. In this case we assume the alternative hypothesis, i.e., that their associated pWCET estimates are not reliable. Since no exceedance was observed for fittings 10 to 20, their *p-value* is 1 and we do not reject the null hypothesis (enough unreliability evidence was not found).

TABLE IV
P-VALUES FOR EACH FITTING OF τ_{g2}

#	$pWCET(10^{-10})$	Obs.Ex. (10^8)	AD1	AD2	p-value
1	31269	1479	0.762	0.748	**0**
2	31274	1378	0.383	0.376	**0**
3	31319	703	0.953	0.952	**0**
4	31428	133	0.224	0.228	**0**
5	31527	29	0.436	0.450	$\mathbf{< 10^{-35}}$
6	31567	13	0.892	0.901	$\mathbf{< 10^{-35}}$
7	31655	1	0.313	0.320	**0.010**
8	31665	1	0.923	0.923	**0.010**
9	31674	1	0.935	0.935	**0.010**
10	31710	0	0.601	0.606	1
11	31854	0	0.862	0.863	1
12	31924	0	0.428	0.428	1
13	31952	0	0.619	0.616	1
14	32039	0	0.148	0.148	1
15	32192	0	0.276	0.272	1
16	32244	0	0.734	0.734	1
17	32380	0	0.261	0.270	1
18	32818	0	0.476	0.472	1
19	36723	0	0.204	0.200	1
20	37549	0	0.773	0.765	1

Fittings 7, 8 and 9 are interesting cases. In these cases we observed a single exceedance of $pWCET(10^{-10})$ in a validation sample of 10^8 measurements. There is a probability of only 1% of $pWCET(10^{-10})$ to be exceeded one or more times in a validation sample of 10^8 measurements (*p-value*=0.01), so we must reject the null hypothesis since we adopted a significance level of 5%.

We should notice the importance of the reliability test. Although all 20 fittings for τ_{g2} passed the GOF test, we were able to reject nine of them with a significance level of 5%, indicating that a GOF test alone is not sufficient to assure the reliability of pWCET estimates produced by MBPTA.

Since we generated the execution times of τ_{g2} from a known distribution Gamma(10000,3), we know that the real value of $pWCET^*(10^{-10})$ equals 31948. The test was able to reject all unreliable estimates but fittings 10, 11 and 12. It rejected fitting 9, that is only 274 clock cycles lower than the actual

pWCET value. A larger validation sample should also reject fittings 10, 11 and 12.

Using again the fact that we know the actual measurement distribution, we can compute the true exceedance probability of fitting 10: $31710 = pWCET^*(1.08 \times 10^{-8})$. The critical value for $\epsilon = 10^{-10}$, $n = 10^8$ and $\alpha = 0.05$ is 1. The power of the reliability test in this scenario, considering $\omega = 1.08 \times 10^{-8}$, is 0.6608 (Eq. 1). It means in this scenario the test has a probability of 66.08% of rejecting $pWCET(10^{-10}) = 31710$.

We also applied the reliability test to pWCET estimates with greater exceedance probabilities ϵ, using again a validation sample of 10^8 measurements. Table V shows the dispersion of $pWCET^j(\epsilon)$ regarding τ_{g2} using exceedance probabilities of 10^{-6} and 10^{-8}, together with the respective observed number of exceedances and *p-value* from the reliability test. Since we generated the execution times of τ_{g2} from a known distribution, we know that the real value of $pWCET^*(10^{-6})$ equals 31447. Only fittings 1 to 5 are unreliable and they were indeed rejected by the reliability test. All reliable fittings were approved by the reliability test, that had a perfect match in this case. The actual value of $pWCET^*(10^{-8})$ equals 31714. In this case fittings 1 to 10 are unreliable, and the reliability test rejected fittings 1 to 7, but did not reject fittings 8 to 10. The results corroborate the usefulness of the proposed test, since it detected potentially unreliable estimates in all considered scenarios. It also makes clear the superiority of the proposed test in relation to HWM-based pWCET reliability evaluation approaches, since it takes into account both the size of the validation sample and the target exceedance probability.

TABLE V
P-VALUES FOR SEVERAL EXCEEDANCE PROBABILITIES OF τ_{g2}

#	$pWCET(10^{-6})$			$pWCET(10^{-8})$		
	pWCET	**Obs.Ex.**	**p-value**	**pWCET**	**Obs.Ex.**	**p-value**
1	31235	2381	**0**	31259	1682	**0**
2	31240	2212	**0**	31264	1584	**0**
3	31273	1400	**0**	31305	868	**0**
4	31358	406	**0**	31405	205	**0**
5	31433	127	**0.005**	31494	49	**0**
6	31450	97	0.631	31524	31	**0**
7	31495	47	1	31598	5	**0.004**
8	31531	28	1	31611	3	0.080
9	31533	26	1	31621	2	0.264
10	31550	19	1	31649	1	0.632
11	31630	2	1	31763	0	1
12	31665	1	1	31823	0	1
13	31675	1	1	31830	0	1
14	31743	0	1	31916	0	1
15	31806	0	1	32024	0	1
16	31831	0	1	32062	0	1
17	31923	0	1	32178	0	1
18	32036	0	1	32438	0	1
19	33117	0	1	34631	0	1
20	33124	0	1	34878	0	1

Since we know the actual distribution of the execution time of the synthetic tasks, we can plot the complementary cumulative distribution function (CCDF or $1-CDF$) of its execution time together with the CCDF of the 20 GEV functions obtained via EVT for the different fittings. Fig. 2 shows such plot for task τ_{g2}. We can see that some of the fitted model curves are optimistic in relation to the actual curve (those that are below), while others are clearly pessimistic (those that are above). The plot highlights the intrinsic variability of probabilistic models produced by EVT, and consequently also clarifies the importance of the proposed test for assessing the execution time bounds produced using them.

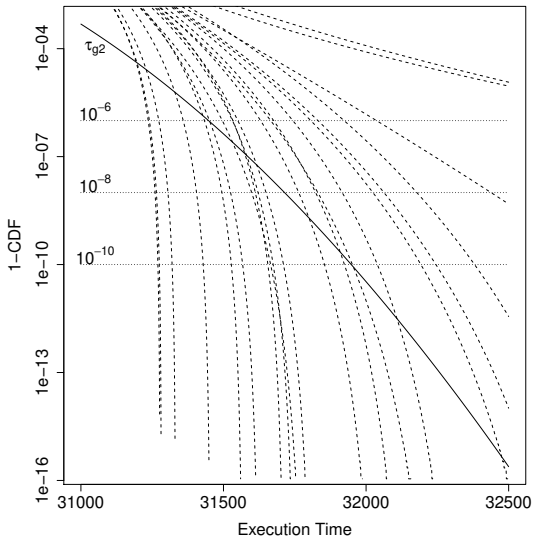


Fig. 2. 1-CDF of task τ_{g2} and several GEV fittings

We repeated the experiments for all five synthetic tasks described earlier. Table VI shows the dispersion of $pWCET^j(10^{-10})$, obtained from 20 different samples of τ_p. Values of pWCET are presented in ascending order for ease of reading, but they could be obtained in any order. For each fitting of τ_p it shows the $pWCET(10^{-10})$, the number e of measurements in the validation sample that exceed each pWCET estimate, the *p-values* obtained from the two versions (AD1 and AD2) of the Anderson-Darling test, and finally the *p-value* obtained when applying the reliability test.

Since we generated the execution times of τ_p from a known distribution Poisson(10000), we know that the real value of $pWCET^*(10^{-10})$ is 10643 and that therefore fittings 1 to 13 are indeed unreliable. Among these the proposed reliability test was able to reject fittings 1 to 7, while all fittings that are in fact reliable were not rejected by the test.

Table VII shows the pWCET estimate, the number of observed exceedances and the respective *p-value* for estimations regarding τ_p using exceedance probabilities of 10^{-6} and 10^{-8}. Since we generated the execution times of τ_p, we know that the real value of $pWCET^*(10^{-6})$ equals 10479. Only fittings 1 to 6 are unreliable and they were indeed rejected by the reliability test, and no reliable fitting was rejected. The actual value of $pWCET^*(10^{-8})$ equals 10566. In this case fittings 1 to 10 are unreliable, and the reliability test rejected fittings 1 to 7, but did not reject fittings 8 to 10.

Table VIII shows the dispersion of $pWCET^j(10^{-10})$, obtained from 20 different samples of τ_{n2}. Values of pWCET are presented in ascending order for ease of reading. For each

fitting of τ_{n2} it shows the $pWCET(10^{-10})$, the number e of measurements in the validation sample that exceed each pWCET estimate, the *p-values* obtained from the two versions (AD1 and AD2) of the Anderson-Darling test, and finally the *p-value* obtained when applying the reliability test.

Since we generated the execution times of τ_{n2} from a known distribution Normal(10000,1000), we know that the real value of $pWCET^*(10^{-10})$ equals 16361. Fittings 1 to 11 are actually unreliable and the reliability test indeed rejected fittings 1 to 10. Although being unreliable, fitting 11 was not rejected. No reliable fitting was rejected by the test.

Table IX shows the pWCET estimate, the number of observed exceedances and the respective *p-value* for estimations regarding τ_{n2} using exceedance probabilities of 10^{-6} and 10^{-8}. The execution times of τ_{n2} were generated from a known distribution so we know that the actual value of $pWCET^*(10^{-6})$ equals 14753. Fittings 1 to 7 are in fact unreliable and they were all rejected by the proposed reliability test, and no reliable fitting was rejected by the test. The actual value of $pWCET^*(10^{-8})$ is 15612. In this case fittings 1 to 10 are unreliable, and the reliability test rejected fittings 1 to 9, but did not reject fitting 10.

TABLE VI
pWCET ESTIMATES, EXCEEDANCE AND P-VALUES FOR τ_p

#	$pWCET(10^{-10})$	Obs.Ex. (10^8)	AD1	AD2	p-value
1	10379	8061	0.072	0.065	**0**
2	10381	7470	**0.034**	**0.030**	**0**
3	10399	3633	0.400	0.388	**0**
4	10440	613	0.322	0.305	**0**
5	10440	613	0.845	0.842	**0**
6	10442	563	**0.020**	**0.018**	**0**
7	10530	9	0.070	0.065	$< 10^{-24}$
8	10563	0	0.220	0.210	1
9	10569	0	**0.037**	**0.033**	1
10	10581	0	0.096	0.089	1
11	10591	0	0.151	0.141	1
12	10604	0	0.239	0.231	1
13	10641	0	0.874	0.901	1
14	10730	0	0.899	0.923	1
15	10762	0	0.063	0.058	1
16	10789	0	0.986	0.984	1
17	10810	0	0.849	0.840	1
18	10818	0	0.423	0.411	1
19	11524	0	0.074	0.069	1
20	11613	0	0.505	0.538	1

TABLE VII
P-VALUES FOR SEVERAL EXCEEDANCE PROBABILITIES OF τ_p

#	$pWCET(10^{-6})$			$pWCET(10^{-8})$		
	pWCET	Obs.Ex.	p-value	pWCET	Obs.Ex.	p-value
1	10375	9399	**0**	10378	8373	**0**
2	10376	9054	**0**	10380	7764	**0**
3	10392	4802	**0**	10398	3788	**0**
4	10424	1277	**0**	10435	763	**0**
5	10425	1215	**0**	10435	763	**0**
6	10427	1094	**0**	10438	661	**0**
7	10493	44	1	10517	13	**0**
8	10505	21	1	10544	3	0.080
9	10516	13	1	10545	3	0.080
10	10518	12	1	10556	1	0.632
11	10533	8	1	10569	0	1
12	10546	3	1	10582	0	1
13	10559	1	1	10608	0	1
14	10600	0	1	10673	0	1
15	10616	0	1	10697	0	1
16	10625	0	1	10714	0	1
17	10635	0	1	10731	0	1
18	10637	0	1	10733	0	1
19	10865	0	1	11165	0	1
20	10920	0	1	11237	0	1

TABLE VIII
pWCET ESTIMATES, EXCEEDANCE AND P-VALUES FOR τ_{n2}

#	$pWCET(10^{-10})$	Obs.Ex. (10^8)	AD1	AD2	p-value
1	13853	5656	0.196	0.197	**0**
2	13976	3341	0.825	0.833	**0**
3	14253	1047	0.136	0.138	**0**
4	14495	352	0.194	0.197	**0**
5	14650	166	0.568	0.568	**0**
6	14845	62	0.177	0.177	$< 10^{-35}$
7	15093	9	0.710	0.701	$< 10^{-24}$
8	15148	7	0.218	0.222	$< 10^{-18}$
9	15389	2	0.218	0.221	$< 10^{-5}$
10	15627	1	0.453	0.448	**0.010**
11	16036	0	0.611	0.613	1
12	16691	0	0.168	0.168	1
13	16795	0	0.775	0.774	1
14	16964	0	0.748	0.761	1
15	17264	0	0.298	0.300	1
16	17611	0	0.127	0.129	1
17	18909	0	**0.036**	**0.037**	1
18	19188	0	**0.028**	**0.028**	1
19	20142	0	0.136	0.137	1
20	27039	0	0.785	0.794	1

VII. EXAMPLE WITH REAL HARDWARE

In this section we illustrate the application of the proposed reliability test with tasks *bsort*, *matmult*, *fdct* and *fir* from the Mälardalen WCET Benchmarks suite [25], as representatives of scenarios in which real code is executed on hardware.

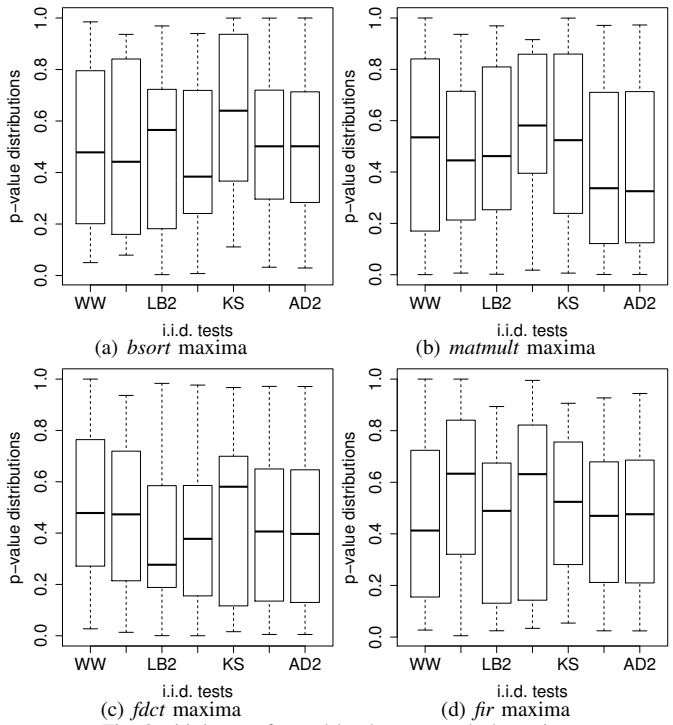
The execution times were obtained from a time-randomized dual-core processor with a simple five-stage pipeline that implements the MIPS instruction set and runs at 50MHz on an FPGA. It employs 512-byte 2-way set-associative private cache memories with modulo placement, write-through update, and randomized replacement policies. It uses an arbitration policy for its separate data and instruction memory buses that determines the next core to be served in a purely random manner. A distinct instance of the measured task is exclusively executed on each core of the processor, without operating system, and measurements used are obtained from the one running at Core#0. Tasks' inputs are fixed, such that a single execution path is exercised and therefore timing variability stems exclusively from the time-randomized hardware. We also configure the processor's ALU to produce maximum (data-independent) latencies during measurements, and fully reset the processor before starting each execution of the tasks. These measures are employed to eliminate dependency between measurements due to hardware state retention.

We collected a modelling sample of 200,000 measurements

TABLE IX
P-VALUES FOR SEVERAL EXCEEDANCE PROBABILITIES OF τ_{n2}

#	$pWCET(10^{-6})$			$pWCET(10^{-8})$		
	pWCET	**Obs.Ex.**	**p-value**	**pWCET**	**Obs.Ex.**	**p-value**
1	13817	6634	**0**	13845	5875	**0**
2	13911	4444	**0**	13959	3570	**0**
3	14153	1620	**0**	14225	1203	**0**
4	14366	630	**0**	14457	421	**0**
5	14446	445	**0**	14585	224	**0**
6	14571	239	**0**	14751	99	**0**
7	14687	140	**0**	14943	30	**0**
8	14841	64	1	15042	12	**0**
9	14853	55	1	15243	5	**0.004**
10	14985	21	1	15299	3	0.080
11	15321	3	1	15751	0	1
12	15655	1	1	16298	0	1
13	15831	0	1	16345	0	1
14	15879	0	1	16502	0	1
15	15956	0	1	16690	0	1
16	16236	0	1	17009	0	1
17	16595	0	1	17800	0	1
18	16874	0	1	18092	0	1
19	17188	0	1	18693	0	1
20	19044	0	1	22574	0	1

and made 20 fittings for each task, using different samples of 10,000 measurements. The Block Maxima approach was used, and a GEV distribution was fitted using the L-moments method [24] to the measurements' maxima selected from blocks of size 100. We checked the applicability of EVT to our execution time measurements using the statistical tests described in Section II. Fig. 3 shows the box-and-whisker plots of those tests applied to tasks *bsort*, *matmult*, *fdct* and *fir*. The *p-values* are acceptable, since they are distributed in the range $[0, 1]$ and do not present any clear tendency to low ($< 5\%$) values.



(a) *bsort* maxima (b) *matmult* maxima (c) *fdct* maxima (d) *fir* maxima

Fig. 3. i.i.d. tests for real-hardware samples' maxima

Throughout this section we consider pWCET estimates with exceedance probability $\epsilon = 10^{-10}$, we use $\alpha = 0.05$, and employ a validation sample with $n = 10^8$ measurements.

Table X shows the estimates of $pWCET^j(10^{-10})$, obtained from 20 different samples of *bsort*. Values of pWCET are presented in ascending order for ease of reading, but they could be obtained in any order. For each fitting of *bsort* it shows the $pWCET(10^{-10})$ estimate, the number e of measurements in the validation sample that exceed each pWCET estimate, and the *p-values* obtained from the two versions (AD1 and AD2) of the Anderson-Darling test used as goodness-of-fit (GOF) test [11]. Using a significance level of 5%, only fittings 15 and 16 would be rejected by the GOF test. We highlight in bold when the *p-value* from the GOF test results in the fitting being rejected at that significance level.

We applied the reliability test described in Section IV to the fittings of task *bsort*. The null hypothesis "pWCET estimate is reliable" is considered to hold only for those fittings that passed the GOF test. Notwithstanding, we applied the reliability test to all 20 fittings. Table X also shows, for each of the 20 fittings of *bsort*, the *p-value* obtained when applying the reliability test. Using a significance level of 5%, the null hypothesis would be rejected for fittings 1 to 11. We highlight in bold when the *p-value* from the reliability test results in the fitting being rejected at that significance level. In those cases we assume the alternative hypothesis, that is, the pWCET estimate is not reliable. Since no exceedance was observed for fittings 12 to 20, their *p-value* is 1 and we do not reject the null hypothesis, i.e., enough unreliability evidence was not found.

TABLE X
P-VALUES FOR EACH FITTING OF *bsort*

#	$pWCET(10^{-10})$	**Obs.Ex. (10^8)**	**AD1**	**AD2**	**p-value**
1	25249	11709	0.275	0.331	**0**
2	25265	1252	0.369	0.364	**0**
3	25267	925	0.114	0.090	**0**
4	25268	786	0.309	0.348	**0**
5	25270	576	0.907	0.958	**0**
6	25270	576	0.340	0.430	**0**
7	25273	361	0.294	0.288	**0**
8	25275	260	0.813	0.828	**0**
9	25280	125	0.070	0.094	**0**
10	25298	3	0.366	0.446	$\mathbf{< 10^{-6}}$
11	25303	2	0.647	0.672	$\mathbf{< 10^{-4}}$
12	25306	0	0.326	0.320	1
13	25309	0	0.487	0.429	1
14	25315	0	0.146	0.177	1
15	25358	0	**0.037**	**0.048**	1
16	25368	0	**0.031**	**0.040**	1
17	25390	0	0.400	0.473	1
18	25392	0	0.587	0.534	1
19	25412	0	0.792	0.808	1
20	25450	0	0.595	0.686	1

This example with *bsort* makes clear the proposed reliability test is complementary to the EVT applicability tests and that both should be used, whenever possible. Considering the 20 fittings done, we would reject fittings 15 and 16 due to the goodness-of-fit test. We should also reject fittings 1 to 11 due to the reliability test. For the sake of safety, only the remaining

fittings (12 to 14 and 17 to 20) should be considered in the development process of an RTS.

Table XI shows the estimates of $pWCET(10^{-10})$ obtained from 20 different samples of *matmult*, the number e of measurements in the validation sample that exceed each pWCET estimate, the *p-values* obtained from the two versions (AD1 and AD2) of the Anderson-Darling test used as goodness-of-fit (GOF) test [11], and the *p-value* produced when applying the proposed reliability test to each of the fittings. Using a significance level of 5%, fittings 5, 8, 11 and 15 would be rejected by the GOF test. The null hypothesis of the proposed reliability test would be rejected for fittings 1 to 13, for which we assume the alternative hypothesis, that is, the pWCET estimates are not reliable. Tables XII and XIII present the same outcomes of the analysis for tasks *fdct* and *fir*, respectively, and both lead to similar conclusions.

Regarding the power of the test, differently from Section VI, in the case of real-hardware tasks we don't know the real value of $pWCET^*(10^{-10})$, naturally. For this reason, we cannot compute the exact power of the test. However, we can rely on Equation 1 in Section V to obtain the power of the test assuming different values for $\omega > 10^{-10}$. Also, Fig. 1 in Section V was built for a sample size of 10^8 and $\alpha = 0.05$ and it shows the power of the test assuming different values for $\omega > 10^{-10}$ and $pWCET(10^{-10}) = pWCET^*(\omega)$.

TABLE XI
pWCET ESTIMATES, EXCEEDANCE AND P-VALUES FOR *matmult*

#	$pWCET(10^{-10})$	Obs.Ex. (10^8)	AD1	AD2	p-value
1	49960	9093	0.423	0.492	**0**
2	49964	5413	0.987	0.998	**0**
3	49968	3151	0.594	0.516	**0**
4	49977	942	0.120	0.154	**0**
5	49987	210	**0.002**	**0.003**	**0**
6	49990	121	0.482	0.603	**0**
7	50000	26	0.855	0.842	$< \mathbf{10^{-35}}$
8	50000	26	**0.004**	**0.005**	$< \mathbf{10^{-35}}$
9	50002	16	0.231	0.259	$< \mathbf{10^{-35}}$
10	50004	12	0.252	0.307	$< \mathbf{10^{-32}}$
11	50013	2	0.063	**0.048**	$< \mathbf{10^{-4}}$
12	50018	1	0.501	0.443	**0.010**
13	50019	1	0.951	0.971	**0.010**
14	50027	0	0.161	0.215	1
15	50032	0	**0.050**	0.067	1
16	50057	0	0.980	0.982	1
17	50060	0	0.843	0.861	1
18	50063	0	0.170	0.224	1
19	50073	0	0.063	0.082	1
20	50242	0	0.075	0.100	1

VIII. Conclusion

In this paper we described a statistical hypothesis test, based on the binomial experiment theory, to test the reliability of pWCET estimates obtained through MBPTA. We illustrated its use with synthetic measurements of hypothetical tasks whose execution times come from EVT-compliant distributions, which is an artificially favourable scenario to MBPTA applicability. These synthetic examples showed the potential value of using a reliability test even when both applicability and GOF tests pass. Then we applied the reliability test to actual hardware measurements, showing how the reliability test effectively complements the MBPTA applicability and GOF tests to detect potentially unreliable pWCET estimates.

TABLE XII
pWCET ESTIMATES, EXCEEDANCE AND P-VALUES FOR *fdct*

#	$pWCET(10^{-10})$	Obs.Ex. (10^8)	AD1	AD2	p-value
1	57442	9124	0.760	0.771	**0**
2	57442	9124	0.371	0.388	**0**
3	57461	4050	0.197	0.210	**0**
4	57469	2781	0.183	0.195	**0**
5	57475	2113	0.217	0.204	**0**
6	57481	1633	0.163	0.173	**0**
7	57495	846	0.977	0.986	**0**
8	57500	679	0.967	0.964	**0**
9	57504	562	0.288	0.272	**0**
10	57524	203	0.959	0.974	**0**
11	57542	87	0.172	0.181	$< \mathbf{10^{-35}}$
12	57589	7	0.061	0.067	$< \mathbf{10^{-17}}$
13	57593	6	0.427	0.411	$< \mathbf{10^{-14}}$
14	57618	3	0.998	0.999	$< \mathbf{10^{-6}}$
15	57644	2	0.477	0.505	$< \mathbf{10^{-4}}$
16	57672	0	0.732	0.761	1
17	57703	0	0.733	0.750	1
18	57740	0	0.135	0.126	1
19	57807	0	0.547	0.536	1
20	57836	0	**0.014**	**0.016**	1

TABLE XIII
pWCET ESTIMATES, EXCEEDANCE AND P-VALUES FOR *fir*

#	$pWCET(10^{-10})$	Obs.Ex. (10^8)	AD1	AD2	p-value
1	50508	4005	0.097	0.106	**0**
2	50551	887	0.356	0.368	**0**
3	50560	631	0.696	0.712	**0**
4	50588	210	0.394	0.403	**0**
5	50616	68	0.336	0.353	$< \mathbf{10^{-35}}$
6	50623	46	0.359	0.345	$< \mathbf{10^{-35}}$
7	50637	18	0.490	0.525	$< \mathbf{10^{-35}}$
8	50664	7	0.985	0.984	$< \mathbf{10^{-17}}$
9	50686	2	**0.004**	**0.004**	$< \mathbf{10^{-4}}$
10	50691	1	**0.001**	**0.001**	**0.010**
11	50797	0	0.892	0.888	1
12	50843	0	0.883	0.888	1
13	50942	0	0.970	0.972	1
14	51016	0	0.725	0.754	1
15	51107	0	0.405	0.429	1
16	51186	0	0.134	0.127	1
17	51331	0	0.900	0.886	1
18	52232	0	0.336	0.354	1
19	52597	0	**0.027**	**0.024**	1
20	53409	0	0.442	0.436	1

EVT-based MBPTA is subject to uncertainty with respect to the estimated parameters of the probabilistic models it employs. This is intrinsic to any method that operates on data samples with significant variability, which is undoubtedly the case for MBPTA. For this reason, any result yielded by MBPTA carries a certain amount of error (in the statistical sense), which can cause it to produce either reliable/pessimistic or unreliable/optimistic pWCET estimates. The paper's contribution is a hypothesis test formulation for testing the reliability of pWCET estimates under such conditions, which can be used for

increasing the confidence that the use of unreliable/optimistic pWCET estimates is avoided.

The experiments presented in Section VI showed, using synthetic measurements from hypothetical tasks with known pWCETs, that the proposed test is capable of detecting unreliable estimates that passed the applicability and GOF tests. For instance, 12 of the 20 $pWCET(10^{-10})$ estimates produced for a task with execution times from a Gamma distribution were known to be in fact unreliable. Although all 20 were approved by the GOF test used, our test rejected 9 of them using a validation sample of 10^8 measurements. For samples with Poisson and Normal distributions, from the 20 fittings produced in the experiments, 13 and 11 of the 20 estimates in each scenario were known to be in fact unreliable, respectively. While only 3 and 2 did not pass the GOF tests used, the proposed reliability test rejected 7 and 10, respectively.

The experiments also showed the proposed test is more effective in accurately detecting unreliable fittings for larger exceedance probabilities. With a validation sample of 10^8 measurements, it was more accurate in detecting unreliability for $pWCET(10^{-8})$ than for $pWCET(10^{-10})$, and produced ideal results for $pWCET(10^{-6})$ estimates. This is expected as explained in Section V regarding the power of the test.

The reliability test presented in this paper uses two counts, only: the number n of task executions and the number e of execution times that exceeded a constant and off-line computed pWCET estimate. It is possible to include two counters per task in the final system, to continuously track the test statistic. The permanent execution of the reliability test could detect whether varying execution conditions through the lifetime of the product affect the reliability of pWCET estimations defined during development. This approach could be used as part of an early fault detection strategy for real-time systems.

There are several open questions regarding MBPTA for real-time systems [5] [6], but the benefits of its use could be enormous. Some MBPTA-based tools are already being developed, such as WOMBAT (Worst-Case Measurement-Based Statistical Tool), that targets avionics applications executed on embedded COTS multi-core processors [26].

For MBPTA based on EVT to be accepted into engineering processes, it is necessary to provide evidence of its reliability. The reliability test based on a binomial experiment described in this paper is a contribution in that direction.

REFERENCES

[1] R. Wilhelm et al., "The worst-case execution time problem—overview of methods and survey of tools," ACM Transactions on Embedded Computing Systems (TECS), vol. 7, p. 36, 04 2008.

[2] L. Cucu-Grosjean et al., "Measurement-based probabilistic timing analysis for multi-path programs," in 2012 24th Euromicro Conference on Real-Time Systems, July 2012, pp. 91–101.

[3] S. Coles, An introduction to statistical modeling of extreme values, ser. Springer Series in Statistics. London: Springer-Verlag, 2001.

[4] F. Reghenzani, G. Massari, W. Fornaciari, and A. Galimberti, "Probabilistic-wcet reliability: On the experimental validation of evt hypotheses," in Proceedings of the International Conference on Omni-Layer Intelligent Systems, ser. COINS '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 229–234. [Online]. Available: https://doi.org/10.1145/3312614.3312660

[5] R. Davis and L. Cucu-Grosjean, "A survey of probabilistic timing analysis techniques for real-time systems," Leibniz Transactions on Embedded Systems, vol. 6, no. 1, p. 60, 2019.

[6] F. J. Cazorla et al., "Probabilistic worst-case timing analysis: Taxonomy and comprehensive survey," ACM Comput. Surv., vol. 52, no. 1, Feb. 2019.

[7] L. Santinelli, F. Guet, and J. Morio, "Revising measurement-based probabilistic timing analysis," in 2017 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS), April 2017, pp. 199–208.

[8] J. Abella et al., "Heart of gold: Making the improbable happen to increase confidence in mbpta," in 2014 26th Euromicro Conference on Real-Time Systems, July 2014, pp. 255–265.

[9] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," Biometrika, vol. 65, no. 2, pp. 297–303, 1978.

[10] L. Santinelli, J. Morio, G. Dufour, and D. Jacquemart, "On the sustainability of the extreme value theory for wcet estimation," in WCET, vol. 39, 07 2014.

[11] F. W. Scholz and M. A. Stephens, "K-sample anderson-darling tests," Journal of the American Statistical Association, vol. 82, no. 399, pp. 918–924, 1987.

[12] M. Lin, H. C. Lucas, and G. Shmueli, "Research commentary—too big to fail: Large samples and the p-value problem," Info. Sys. Research, vol. 24, no. 4, p. 906–917, Dec. 2013. [Online]. Available: https://doi.org/10.1287/isre.2013.0480

[13] I. Haigh and T. Wahl, "Advances in extreme value analysis and application to natural hazards," Natural Hazards, vol. 98, pp. 1–4, 08 2019.

[14] M. Liu, M. Behnam, and T. Nolte, "Applying the peak over thresholds method on worst-case response time analysis of complex real-time systems," in 2013 IEEE 19th International Conference on Embedded and Real-Time Computing Systems and Applications, 2013, pp. 22–31.

[15] B. Lesage et al., "A framework for the evaluation of measurement-based timing analyses," in Proceedings of the 23rd International Conference on Real Time and Networks Systems, ser. RTNS '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 35–44. [Online]. Available: https://doi.org/10.1145/2834848.2834858

[16] J. Abella, M. Padilla, J. D. Castillo, and F. J. Cazorla, "Measurement-based worst-case execution time estimation using the coefficient of variation," ACM Trans. Des. Autom. Electron. Syst., vol. 22, no. 4, Jun. 2017. [Online]. Available: https://doi.org/10.1145/3065924

[17] K. P. Silva, L. F. Arcaro, and R. Silva De Oliveira, "On using gev or gumbel models when applying evt for probabilistic wcet estimation," in 2017 IEEE Real-Time Systems Symposium (RTSS), Dec 2017, pp. 220–230.

[18] L. Arcaro, K. Silva, and R. Oliveira, "On the reliability and tightness of gp and exponential models for probabilistic wcet estimation," ACM Transactions on Design Automation of Electronic Systems, vol. 23, pp. 1–27, 03 2018.

[19] L. F. Arcaro, K. Palma Silva, and R. Silva De Oliveira, "A reliability evaluation method for probabilistic wcet estimates based on the comparison of empirical exceedance densities," in 2018 VIII Brazilian Symposium on Computing Systems Engineering (SBESC), Nov 2018, pp. 196–200.

[20] F. Reghenzani, G. Massari, L. Santinelli, and W. Fornaciari, "Statistical power estimation dataset for external validation gof tests on evt distribution," Data in Brief, vol. 25, p. 104071, 2019.

[21] M. Fernandez et al., "Probabilistic timing analysis on time-randomized platforms for the space domain," in Design, Automation & Test in Europe Conference Exhibition (DATE), 2017, 2017, pp. 738–739.

[22] D. Montgomery and G. Runger, Applied statistics and probability for engineers, 7th ed. Wiley, 2018.

[23] R. R. Holmes Jr. and K. Dinicola, "100-Year Flood – It's All About Chance," U.S. Geological Survey General Information Product, vol. 106, p. 1, 2010.

[24] J. R. M. Hosking, "L-moments: Analysis and estimation of distributions using linear combinations of order statistics," Journal of the Royal Statistical Society. Series B (Methodological), vol. 52, no. 1, pp. 105–124, 1990. [Online]. Available: http://www.jstor.org/stable/2345653

[25] J. Gustafsson, A. Betts, A. Ermedahl, and B. Lisper, "The mälardalen wcet benchmarks: Past, present and future." in WCET, ser. OASICS, B. Lisper, Ed., vol. 15. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, 2010, pp. 136–146.

[26] P. G. Zaykov and J. Kubalčik, "Worst-case measurement-based statistical tool," in 2019 IEEE Aerospace Conference, 2019, pp. 1–10.