On the Representativity of Execution Time Measurements: Studying Dependence and Multi-Mode Tasks

Fabrice Guet[1], **Luca Santinelli**[2], **and Jerome Morio**[3]

1 **ONERA Toulouse, Toulouse, France**
2 **ONERA Toulouse, Toulouse, France**
3 **ONERA Toulouse, Toulouse, France**

Abstract

The Measurement-Based Probabilistic Timing Analysis (MBPTA) infers probabilistic Worst-Case Execution Time (pWCET) estimates from measurements of tasks execution times; the Extreme Value Theory (EVT) is the statistical tool that MBPTA applies for inferring worst-cases from observations/measurements of the actual task behavior. MBPTA and EVT capability of estimating safe/pessimistic pWCET rely on the quality of the measurements; in particular, execution time measurements have to be representative of the actual system execution conditions and have to cover multiple possible execution conditions. In this work, we investigate statistical dependences between execution time measurements and tasks with multiple runtime operational modes. In the first case, we outline the effects of dependences on the EVT applicability as well as on the quality of the pWCET estimates. In the second case, we propose the best approaches to account for the different task execution modes and guaranteeing safe pWCET estimates that cover them all. The solutions proposed are validated with test cases.

1998 ACM Subject Classification C.3 Real-Time and Embedded Systems

Keywords and phrases Measurement-Based Probabilistic Timing Analysis, probabilistic Worst-Case Execution Time, Extreme Value Theory, Execution Time Measurements Representativity

Digital Object Identifier 10.4230/OASIcs.WCET.2017.3

1 Introduction

Multi-core and many-core processors are becoming common implementations for real-time systems. The large amount of available resources allows increasing performance and embedding multiple functionalities within systems. However, real-time modeling and analysis become more complex due to the increased source of unpredictabilities, [11, 22]; for instance, tasks execution time exhibits variabilities from the runtime dependence/interference between system elements which are difficult to model accurately e.g., access to shared mameory, [21].

Probabilistic timing analysis approaches are being proposed to cope with real-time system unpredictabilities. They consider both the task average execution behavior and the worst-case execution behavior as random variables. In particular, the probabilistic Worst-Case Execution Time (pWCET) extends the notion of Worst-Case Execution Time (WCET) as the worst-case distribution that upper bounds the task execution times. pWCET models have multiple values, each with an associated probability of being the task worst-case execution time; very unlikely cases such as faults are included. This makes pWCET task models more flexible and potentially less pessimistic than classical deterministic WCET (either statically or measurement-based driven) in representing the task behavior. Figure 1 gives an example

17th International Workshop on Worst-Case Execution Time Analysis (WCET 2017).
Editor: Jan Reineke; Article No. 3; pp. 3:1–3:13
Open Access Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

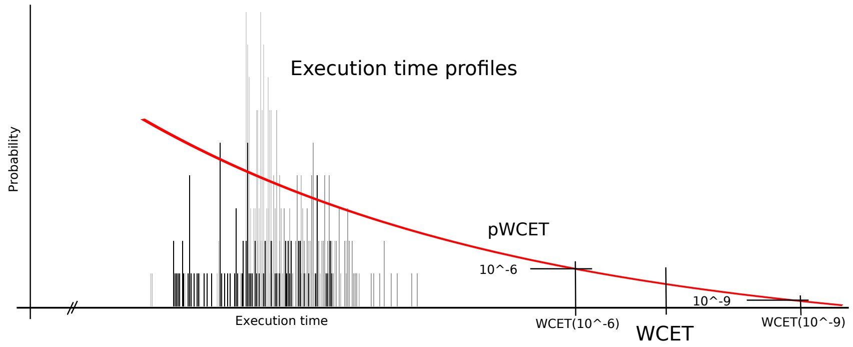

Figure 1 WCET and pWCET representations.

of pWCET and WCET representations to the task behavior; both upper bound the task actual execution times which could be different execution time profiles depending on the system execution conditions.

Measurement-Based Probabilistic Timing Analysis. The Measurement-Based Probabilistic Timing Analysis (MBPTA) is a probabilistic timing analysis that makes use of measurements of task execution times for computing pWCET estimates. The Extreme Value Theory (EVT) applied in the MBPTA allows for inferring the rare events (worst-case bounds) from observations of the actual task behavior (measurements). MBPTA does not need accurate system nor task models, instead they demand measurements of execution time representative of all the system execution behaviors.

A first application of the EVT for the timing analysis of real-time systems considers Gumbel distributions for the pWCET estimates, [4]. In [6, 1, 5], only artificially[1] time randomized real-time systems are analyzed with the EVT. Last developments in MBPTA propose a generalized version of the EVT [18, 12] which can be applied to both non-time randomized real-time systems and artificially time randomized real-time systems.

MBPTA Open Problems. Today's MBPTA works have completely defined the EVT and its applicability to the pWCET problem. The hypotheses for applying the EVT have been deeply investigated, and the quality (as safety and accuracy) of the resulting pWCET estimates has reached good levels. Actual MBPTA challenges are moving to the *representativity* of the execution time measurements, since in order to let the EVT be able to estimate safe worst-case distributions[2], the measurements have to be "good representation" of the system behavior,[20]. We hereby consider a notion of measurement representativity as the capability of capturing any event that characterizes the current system behavior. Those events would be dependence between consecutive executions, pattern of executions e.g., cyclic execution times or clusters, multiple execution conditions or operational modes, etc..

In [1], the system is artificially randomized in order to make the appearance of worst-case measurements more probable. For those systems, some works have approached the problem of measurement representativity [2, 20] where the representativity notion considered restricted

[1] By artificially time randomized systems we mean systems where there have been added randomization mechanisms such as random replacement caches or random task re-mapping in memory at each execution.

[2] Safe worst-case execution time distributions (pWCETs) are distributions that upper bound any possible task execution.

to the capability of measurements of capturing worst-case events. instead, we hereby consider a broader notion of representativity that includes the capability of capturing dependence or pattern of executions.

Full coverage of tasks/system input conditions and measurements, which needs to include pathological cases and their large execution times, have to be guaranteed to MBPTA. As today, they remain open problems related to the representativity of the measurements of any architecture, including artificial randomized real-time systems.

In this paper we focus on two aspects of the representativity which real-time systems face constantly: statistical dependence between measurements and tasks with multiple operational modes.

We intend to demonstrate that measurements which are representative of the dependent system behavior have to be preserved and not modified whatsoever. Instead, with measurements which are representative of multiple execution condition, worst-case behaviors or modes cannot be neglected.

- The statistical dependence between measurements is when from one set of measurements it is possible to infer future measurements e.g., clusters of measurements or consecutive measurements with similar values appearing periodically. With real-time systems, examples of dependences are series of specific execution conditions e.g., bursts of interferences or cache locality, or task inputs/execution conditions that appear periodically. *Would it possible to apply the EVT in case of dependence between measurements? What is the impact that dependences have on pWCET estimates?*
- Real-time tasks can be implemented with multiple operational modes e.g., taking-off, cruising and landing modes which alternate at runtime in avionic systems. More simple examples are multi-path tasks where, depending on the input applied a path can be triggered with consequently different execution time. *How is it possible to apply the EVT to multi-mode execution time measurements? What is the pWCET estimate that guarantees all the modes, a.k.a. a safe pWCET estimates?*

If the system shows dependent behavior and multi-mode tasks, the measurements have to embed such events in order to characterize the system behavior and being representative for it. To the previous questions we provide answers with this work.

Contributions: We propose guidelines for letting EVT and MBPTA tackle with dependent measurements of execution times and multi-mode real-time tasks. We describe what has to be done in both cases in order to correctly apply the EVT and obtain safe and accurate pWCET estimates. We provide also a statistical analysis to the measurements for identifying the limits of the EVT application and the conditions for qualitatively defining the representativity of the measurements in terms of dependence and multi-mode tasks. This would allow to extend EVT applicability to more realistic real-time systems e.g., with statistical dependence or multi-path tasks. Our contributions are validated with test cases from industrial applications, multi- and many-core real-time systems and artificial traces of execution time measurements.

Organization of the paper: In Section 2 we state some background for the MBPTA, the EVT and the probabilistic modeling of average and worst-case task execution behavior. Section 3 details the EVT applicability in case of statistical dependent measurements and the effects that dependence has on pWCET estimates; case studies are applied to validate the guarantees that EVT offers to task pWCETs in case of dependence. Section 4 approaches the challenge of EVT applicability to multi-mode tasks; solutions to guarantee pWCET estimates with multiple execution conditions are developed and validated with case studies. Section 5 is for conclusions and future work.

2 Background: Probabilistic Modeling and Extreme Value Theory

A trace $\mathcal{T}$ is a collection of execution time measurements C_j, $\mathcal{T} = \{C_j \mid j \in [\![1:n]\!]\}$, n is the size of the trace. Given $\mathcal{T}$, the Execution Time Profile (ETP) $\mathcal{C}$ is the discrete random variable defined on the finite support $\Omega_{\mathcal{C}}$ of possible execution time values $C_{(k)}$, $\Omega_{\mathcal{C}} = (C_{(k)})_{k\in[\![1:N]\!]}$ with $C_{(k)} \in \mathcal{T}$; N is the number of different value in $\mathcal{C}$. The ETP is an empirical discrete random variable[3] that describes the task actual execution behavior. Representations for $\mathcal{C}$ are: the discrete probability mass function or Probability Distribution Function (PDF) $\mathsf{pdf}_{\mathcal{C}}$, such that $\mathsf{pdf}_{\mathcal{C}}(C_{(k)}) = P(\mathcal{C} = C_{(k)})$, the empirical Cumulative Distribution Function (CDF) $\mathsf{cdf}_{\mathcal{C}}$ as the discrete function $\mathsf{cdf}_{\mathcal{C}}(C) = P(\mathcal{C} \leq C)$ with $\mathsf{cdf}_{\mathcal{C}}(C) \in [0;1]$ and the Complementary Cumulative Distribution Function (CCDF) $\mathsf{icdf}_{\mathcal{C}}$ defined by the probability of exceeding the execution time threshold C (risk probability), $\mathsf{icdf}_{\mathcal{C}}(C) = P(\mathcal{C} > C) = 1 - \mathsf{cdf}_{\mathcal{C}}(C)$ with $\mathsf{icdf}_{\mathcal{C}}(C) \in [0;1]$.

Probabilistic timing analysis approaches look for pWCET distribution estimates $\overline{\mathcal{C}}$ that upper bounds any possible task execution behavior. Representations for $\overline{\mathcal{C}}$ are the PDF $\mathsf{pdf}_{\overline{\mathcal{C}}}(C)$ either continuous or discrete, the CDF $\mathsf{cdf}_{\overline{\mathcal{C}}}(C)$ and the CCDF $\mathsf{icdf}_{\overline{\mathcal{C}}}(C)$. $\overline{\mathcal{C}}$ has to be a *safe/pessimistic* representation of the task worst-case behavior: it has to be larger than or equal to[4] any ETP $\mathcal{C}^j$ the task can have, $\mathsf{icdf}_{\overline{\mathcal{C}}}(c) \geq \mathsf{icdf}_{\mathcal{C}^j}(c)$ for every c and every execution condition j. $\overline{\mathcal{C}}$ has also to be a *tight* upper bound to the ETPs; the tightness is for the quality of the pWCET estimates.

2.1 The Extreme Value Theory in a Nutshell

The EVT applies with the Block Maxima (BM) paradigm or with the Peak over Threshold (PoT) paradigm. The BM EVT models the limit law of execution time maxima of blocks of execution time measurements; the PoT EVT models the limit law of the execution times greater than a threshold (peaks above the threshold).

At the infinite (infinite number of block maxima or infinite number of peaks over the threshold) the law of extreme measurements tends to a Generalized Extreme Distribution (GED) or a Generalized Pareto Distribution (GPD) if and only if: i) the input $\mathcal{T}_{\mathcal{C}}$ is composed of independent and identically distributed (iid) measurements and ii) the resulting distribution $\mathcal{C}$ belongs to the Maximum Domain of Attraction (MDA) of the limit distribution. GED and GPD are the limit distribution respectively for the BM and the PoT EVT.

The identical distribution hypothesis assumes that all the measurements C_j follow the same distribution $\mathcal{C}$. The independence hypothesis (statistical independence) assumes that the individual execution time measurements $C_1, \ldots, C_n$ are not correlated with each other. The MDA hypothesis (also named matching) seeks if the limit law of the input distribution $\mathcal{C}$ converges to a GEV or a GPD. The limit law from the EVT is the pWCET estimates $\overline{\mathcal{C}}$.

Generalized EVT. Recent works prove the EVT applicability with more relaxed hypotheses than iid and MDA, [17, 12, 18]. They formalize the so called *generalized EVT* or practical EVT,

[3] $\mathcal{C}$ is a discrete distribution since execution time C_j can only assume values multiple of the system tick. Calligraphic letters are for both random variables, discrete or continuous, and traces, $\mathcal{C}$ and $\mathcal{T}$. Non-calligraphic letters are for single value variables C_j and $C_{(k)}$.

[4] The partial ordering between distribution is defined according to [8].

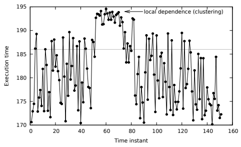

Figure 2 Local dependence between execution time measurements.

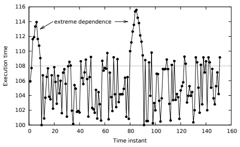

Figure 3 Extreme dependence between execution time measurements.

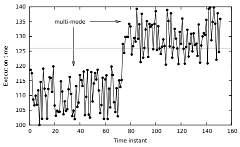

Figure 4 Example of two-mode task.

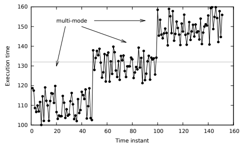

Figure 5 Example of three-mode task.

since it applies to practical cases of real-time systems e.g., not infinite measurements, system without artificially time randomization. The generalized EVT relies on: the stationarity hypothesis h'_1, the short range independence (negation of short range dependence) hypothesis $h'_{2.1}$, the extremal independence (long range dependence) hypothesis $h'_{2.2}$ and the matching hypothesis h'_3. If the EVT follows all the hypothesis, then it provides a safe estimation of the extreme execution times of $\mathcal{C}$.

The *stationarity hypothesis* h'_1 tests if the measurements are stationary and follow the same distribution i.e. the identical distribution. The Kwiatowski Philips Schmidt Shin (KPSS) test [13] checks if the trace is stationarity.

The dependence between measurements instantiates into local dependence i.e. close dependent measurements in $\mathcal{T}_\mathcal{C}$, and dependence between extreme measurements i.e. far measurements.

The *short range dependence (or local dependence)* $h'_{2.1}$ focuses on the relationship between measurements close-within-$\mathcal{T}_\mathcal{C}$. Condition D in [16] formalizes the minimum degree of short range dependence for the EVT applicability; it ensures that for distant enough dependent measurements (short range dependence), the limit law of the peaks over a threshold is still a GPD. A valuable test for the short range dependence $h'_{2.1}$ is the Brock Dechert Scheinkman (BDS) test [3]. Figure 2 gives an example of local dependence between measurements of execution times: execution times above a threshold can cluster as a result of dependence and consecutive interferences between system elements.

The *long range dependence (or extremal dependence)* $h'_{2.2}$ focuses on the relationship between far-in-time measurements. Condition D' in [16] formalizes the minimum degree of long range dependence for the EVT applicability. The extremal index θ, $\theta \in]0;1]$ [9], indicates the degree of clustering of either the PoT or the BM. θ expresses the probability of having distant enough measurements which are independent: the more the peaks or the maxima are distant from each other the more the independence is, and the higher is the probability of having independence it is. θ, with one of its estimators is applied to verify the extremal independence $h'_{2.2}$ [12, 10]. Figure 3 gives an example of extremal dependence between measurements of execution times; patterns that repeat are impacted one another.

The *matching hypothesis* h'_3 is for verifying that $\mathcal{C}$ belongs to the MDA of the GPD. A good matching test is the Cramer Von Mises criterion (CVM) which measures the distance between the empirical CDF of the extreme measurements and the $\overline{\mathcal{C}}$ estimated. The CVM test verifies the validity of h'_3 and it has been chosen because it performs well in the case of extreme value distributions [14].

DIAGXTRM [12] is a MBPTA tool that implements the generalized EVT and we use it to investigate statistical dependence and multi-mode tasks in this work. It implements the PoT EVT version as well as the tests previously described. It also defines confidence levels cl_i to verify the confidence on the EVT applicability hypotheses, thus the confidence on the pWCET estimates; cl_i defines the confidence level on $h'_i \in H'$. cl_i is an integer value, $cl_i \in \mathcal{N}$, and defined in $[0,4]$, $cl_i \in [0,4]$, such that for $cl_i = 0$ there is no confidence in accepting h'_i; for $cl_i = 1$ there is moderate confidence in accepting h'_i; for $cl_i = 2$ there is good confidence in accepting h'_i and so until level 4 with the maximum confidence. The cls are represented with radar plots. Hypothesis testing and other statistics are applied by diagXtrm to evaluate execution time patterns and other characteristics that traces can exhibit. diagXtrm is available at `https://forge.onera.fr/projects/diagxtrm2`.

3 EVT & Dependences

In this section we detail the effects that statistical dependence of measurements has on pWCET estimates.

The pWCET estimate in case of extremal independence $\overline{\mathcal{C}}^{ei}$ is greater than or equal to The pWCET estimate in case of independence $\overline{\mathcal{C}}^{i}$: $\mathsf{icdf}_{\overline{\mathcal{C}}^{ei}} \geq \mathsf{icdf}_{\overline{\mathcal{C}}^{i}}$, [10]. The partial ordering between $\overline{\mathcal{C}}^{ei}$ and $\overline{\mathcal{C}}^{i}$ is ensured if and only if both $\overline{\mathcal{C}}^{ei}$ and $\overline{\mathcal{C}}^{i}$ follow the same average distribution $\mathcal{C}$.

The inequality states that dependence, up to a certain degree which assure EVT applicability (local dependence and extreme dependence $h_{2,1} \wedge h_{2,2}$), provides more pessimistic pWCET estimates than the independence. In other words, clusters of execution times within the trace (short range dependences) or patterns between far away measurements (extremal dependences) make rare events more probable. The EVT accounts for that with more pessimistic pWCET i.e. more conservative pWCET estimates, and the pWCET with dependence remains a safe modeling of the task worst-case.

We remark that the former condition poses some issues to works that aim at creating independence between measurements, [19, 10]. Techniques like re-sampling and de-clustering that are normally applied in some domains, have to be thoroughly investigated before being applied to the pWCET problem because they can produce optimistic pWCETs.

Trace $\mathcal{T}_{\overline{\mathcal{C}}}$ cannot be changed if we aim at guaranteeing safe pWCET. The representativity of the measurements in characterizing the actual task behavior with dependence effects cannot be modified. Eventual changes to assure full statistical independence between measurements and "better-to-apply-EVT" could produce optimistic pWCETs. The lost of representativity of the measurements could end up into unsafe pWCET estimates.

Dependence Case Study. Some test cases are applied for validating the dependence impact on pWCET estimates we propose; each test case is represented by a trace of execution time measurements. The execution times which are measured with the tools available for the test cases, are all in CPU cycles.

- *trace1* is a trace of execution time measurements from an industrial avionic safety-critical multi-core system, [24]; the task under observation executes on one core while another core is doing interfering I/O activities. In its original version (V1), *trace1* has weak extreme independence; from V1 we obtain a modified version of *trace1*, V2 by randomly re-sampling the measurements; V2 has stronger extreme independence than V1.
- *trace2* is a trace of execution time from the *dijkstra* task of the TACLeBench[5] executing on a Kalray many-core platform, [21]. The task under observation executes on one core, while other cores produce interference through shared memory. In its original version V1,

[5] `http://www.tacle.eu/index.php/activities/taclebench`.

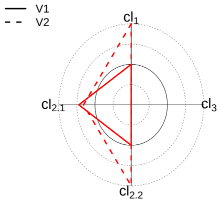

Figure 6 Radar plot for *trace1* with two degrees of extreme independence applied.

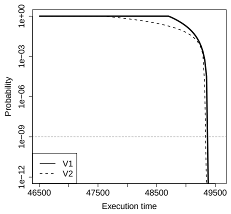

Figure 7 CCDF representation of pW-CETs from PoT EVT applied to *trace1*.

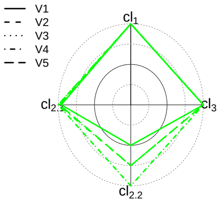

Figure 8 Radar plot of *trace2*: pW-CET confidence of the 5 versions for *trace2*.

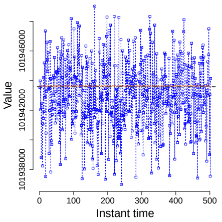

Figure 9 Portion of trace of execution time measurements for *trace2* V1.

trace2 has weak extreme independence. We define 4 more versions of *trace2* by randomly re-sampling the measurements; intermediate versions V2, V3 and V4 have decreasing degree of extreme dependence (more extreme independence); V5 has full independence between measurements.

- *trace3* and *trace4* are artificial traces of execution time measurements extracted from a Gaussian distribution. To *trace3* there has been added local dependence between measurements to reproduce the effect of non-time randomized system elements to the task execution behavior e.g., locality effects from caches. To *trace4* there has been added extremal dependence between measurements to reproduce the effect of periodic inputs. *trace3* has been modified from its original version V1 with local dependence with 4 more versions. V5 has full independence obtained by randomly re-sampling measurements in *trace3* while intermediate versions V2, V3 and V4 have decreasing degree of dependence. *trace4* has been modified from its original version V1 with extreme dependence by randomly re-sampling the measurements; V2 is the least re-sampled version of *trace4* with still strong dependence, V3 is more re-sampled than V2 and V4 is an even more re-sampled version of *trace4*.

The proposed case study is a small fraction of the benchmarks investigated; it is representative because the composing traces exhibit dependence and through which it is possible to illustrate the effects of artificially induced independence on the safety of the pWCETs. The traces are processed with DIAGXTRM for deriving safe and confident pWCETs; all of them and more, except the industrial one, are available at `https://forge.onera.fr/projects/diagxtrm2`.

Dependence Results. Figure 6 details the confidence level of *trace1* in its original version V1 and for the modified version V2; $cl_{2.2}$ increases from V1 to V2 as result of the artificially induced independence on the measurements. Figure 7 illustrates the effect of extreme independence on the pWCET estimates: *by increasing the independence the pWCET estimate decreases.*

To *trace1* the EVT is not applicable i.e. the matching hypothesis fails with $cl_3 = 0$; nonetheless, it helps us to understand the impact of extremal independence on pWCET estimates. Artificially inducted independence, even if with small effects as for *trace1*, acts reducing the pWCET estimates. *If the representativity of a system with dependence is not guaranteed in case of dependent measurements, the risk is to have unsafe pWCET estimates.*

Figure 8 details the radar plot for the EVT approach applied to *trace2*. The confidence levels for the 5 versions of *trace2* are represented and tells that the EVT is confidently applicable to all the versions, included those with certain degree of dependence; $cl_{2.2}$ increases from V1 to V5, meaning that the random re-sampling applied is able to break the extreme

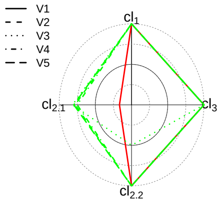

Figure 10 Radar plot of *trace3*: confidence levels for the 5 versions of *trace3*.

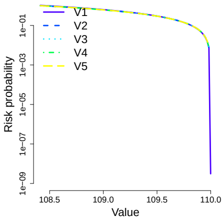

Figure 11 CCDF representation of pWCETs from PoT EVT applied to *trace3* and its modified versions.

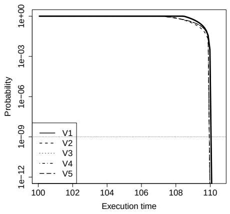

Figure 12 CCDF representation of pWCETs from PoT EVT applied to *trace3* and its modified versions.

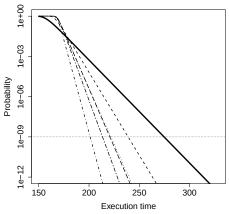

Figure 13 CCDF representation of pWCETs from BM EVT applied with increasing block size to *trace3*.

dependence existing within *trace2* V1. For *trace2*, the pWCET variations due to artificial induced independence are negligibly small with respect to the dependent case V1: there is no impact on the pWCET of extreme dependence estimates for *trace2*. It is a particular case, due to the trace small variability and the shape of the resulting pWCET.

Figure 9 illustrates a portion of the trace of measurements for *trace2* V1; although no particular input or execution condition are exercised (dijkstra task in a many-core execution with interference generated from a concurrent task with no particular input imposed, [21]), the behavior of the peaks over the threshold selected (horizontal line) could follow a oscillatory periodic pattern, hence some degree of extreme dependence exists. Future work will focus on defining possible patterns with statistics.

Figure 11 details the PoT EVT approach applied to *trace3*. The pWCET variations of the version V2, V3, V4 and V5 are small with respect to the dependent case V1, but they all act reducing the pWCET estimates. The guarantee of having safe pWCETs from artificially independent traces reduces because the trace modifications are not conservative. Figure 10 illustrates the radar plot of the confidence levels for the 5 versions of *trace3*: the EVT is confidently applicable to all the versions, including those with certain degree of dependence.

We hereby specialize the comparison between PoT and BM, proposed first in [23], to the statistical dependence case. The BM EVT is applicable to more dependent cases than the PoT, since grouping consecutive measurements into blocks and selecting only the maximum of each block would break possible local dependences and extremal dependences. We say that the BM EVT is more robust with respect to dependence (local dependence and/or extremal dependences) than the PoT because of the capacity of block maxima of filtering measurements. Also, the PoT filters measurements i.e. those below the threshold, but it does not with respect to dependences. The problem with BM is that filtering dependences with large block sizes would reduce the impact of dependence on the rare events, thus resulting into possibly optimistic pWCET estimates. Figure 13 details the BM EVT applied to *trace4* with different block sizes. The continuous line pWCET is for block size of 10 measurements, the dotted line for block size of 20, and so on. The more the block size increases and the dependence between block maxima decreases, the more the pWCET estimates decreases augmenting the risk of optimistic pWCET estimates.

Figure 12 details the PoT EVT approach applied to *trace4*. The pWCET variations induced by less dependence (versions V2 to V5) are small with respect to V1, but act reducing the pWCET estimates questioning the safety of pWCET estimates with artificial independent traces.

By comparing BM and PoT with respect to dependence effects, we observe that the PoT is a more robust approach than BM to independence effects (artificially created independence).

Nonetheless, with neither of them it is possible to guarantee safe pWCET estimates if the traces loose their representativity with artificially reduced dependence. We are currently employing robustness as intuitive notion, future work will be devoted to formalize it for the EVT.

Forcing independence into dependent execution time measurements makes the trace not representative anymore of the dependent system behavior. As a result, the pWCET could end up into a non safe anymore worst-case estimation of the task execution behavior.

4 EVT & Multi-Mode Tasks

The pWCET estimate depends on the execution conditions applied for the measurements; the EVT is able to produce the worst-case bound for that condition only.

Measurement-based timing analysis, either probabilistic MBPTA or deterministic [15], have to cover every possible execution condition and inputs in order to guarantee the absolute worst-case bound estimate. Thus, *the measurements have to be representative of all the possible execution conditions the system can experience.*

With $J = \{j\}$ the finite set of possible measurement execution conditions for a system, there exist two ways of integrating all the scenarios into the MBPTA:
Trace-merging consists of merging all the traces $\mathcal{T}_{\mathcal{C}^j}$ $\forall j \in J$ within a unique trace $\mathcal{T}_{\mathcal{C}}$, $\mathcal{T}_{\mathcal{C}} \overset{def}{=} \bigcup_{j \in J} \mathcal{T}_{\mathcal{C}^j}$; the EVT is applied to $\mathcal{T}_{\mathcal{C}}$ for deriving $\overline{\mathcal{C}}$ as the worst-case distribution for J.
Envelope consists of applying the EVT to each measurement condition j and get $\overline{\mathcal{C}}^j$ for all $j \in J$. The worst-case distribution $\overline{\mathcal{C}}$ that upper bounds every $j \in J$ is such that: $\overline{\mathcal{C}} \overset{def}{=} max_{j \in J}\{\overline{\mathcal{C}}^j\}$ and $\mathsf{icdf}_{\overline{\mathcal{C}}}(C) \overset{def}{=} max_{j \in J}\{\mathsf{icdf}_{\overline{\mathcal{C}}^j}(C)\}$.

Measurements representativity with respect to the system behaviors needs the measurements to include every possible execution behavior for the task. Instead of enumerating all the possible execution conditions for a system, the dominance between some of them ($\mathsf{icdf}_{\overline{\mathcal{C}}^1} \geq \mathsf{icdf}_{\overline{\mathcal{C}}^2}$) and the knowledge of worst conditions would keep the measurements representative and allow for worst/safe pWCET estimates. Both trace-merging and envelope approaches rely on knowing all the measurement conditions.

Task inputs and operational modes contribute to define the execution conditions of the system and its tasks. Hence, the behavior of multi-mode tasks is assimilated to multiple execution conditions; trace-merging and envelope are the approaches that can be applied to multi-mode tasks for guaranteeing worst pWCET and measurement representativity. Figure 4 presents a trace of execution time measurements for a two-mode task, while Figure 5 presents the three-mode task case.

Multi-Mode Case Study. Some test cases are applied for validating the multi-mode task study we propose; each test case is represented by a trace of execution times measurements. The traces are representative of actual execution conditions; the execution times, which are measured using the tools available for the test cases, are all in CPU cycles.

- *trace5* is a trace from an industrial avionic safety-critical embedded system (different than *trace1*) where the task is a two-mode application executing on a multi-core platform. The task under observation executes on one core, while another core is doing interfering I/O activities, [24];
- *trace6* comes from the *ns* Mälardalen benchmark task implementation[6] executed in

[6] `http://www.mrtc.mdh.se/projects/wcet/benchmarks.html`.

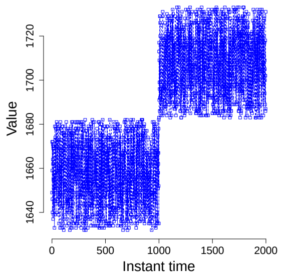

Figure 14 *trace5* with two operational modes.

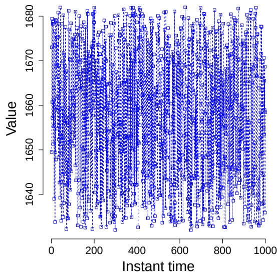

Figure 15 *trace5* with only the first mode represented.

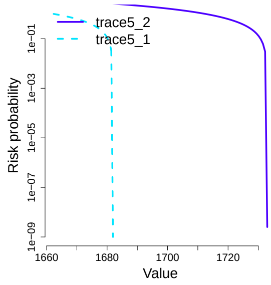

Figure 16 CCDF representation of *trace5* decomposed into *trace5_1* and *trace5_2*.

isolation on a multi-core real-time system, [12]. *trace6* describes a four-mode task where the 4 inputs (for 4 task paths) are randomly picked at runtime;

- *trace7* is a trace from a FPGA implementation of a multi-core real-time system; it is obtained from the *lms* task of the Malardalen benchmark which trace has been obtained by running the task under observation with interference from other tasks, [7]. The task trace results into a two operational mode of execution times not controlled whatsoever (no particular input exercised), and the measurements capture the two different modes with their effects on the execution time.

Among the possible benchmarks investigated, we report few representative cases of multi-mode tasks. The traces are processed with DIAGXTRM; all of these, except the industrial one, are available at `https://forge.onera.fr/projects/diagxtrm2`.

Multi-Mode Results. Figure 14 illustrates the two-mode task represented by *trace5*; Figure 15 details the first part of the trace *trace5_1*, which describe task execution mode 1. In order to apply the EVT, *trace5* has to be decomposed into two traces *trace5_1* and *trace5_2*, respectively for the first mode and for the second mode. Only the envelope approach can be used with *trace5* and the reason is that without decomposing *trace5* into two traces, h'_1 and h'_3 cannot be verified since the peaks above the threshold would belong to both modes. Figure 16 details the pWCETs from the two parts of *trace5*, with *trace5_2* dominating *trace5_1*. For *trace5* if is sufficient to have measurements representative of the worst mode to guarantee the worst pWCET.

The two-mode *trace6* is illustrated in Figure 17; it is possible to apply the EVT to *trace6* with the trace-merging approach, because the peak above the threshold would belong only to the mode with larger execution times (worst mode). With that, both h'_1 and h'_3 can be confidently verified. The envelope and the trace-merging approaches produces same pWCETs for *ns* (*trace6*) because the worst-case condition dominates all the others. Figure 18 depicts the pWCET estimate and the perfect fit between the pWCET distribution and the measured execution time peaks (best fitting the input measurements is a critical element for the EVT application).

The two-mode *trace7* is detailed in Figure 19. In order to apply the EVT, *trace7* has to be decomposed into *trace7_1* and *trace7_2*, respectively for the first mode and for the second mode. Only the envelope approach can be used with *trace7* and the two separated traces. Figure 20 details the pWCETs from the two parts of *trace7*. The worst pWCET is the maximum of the two pWCET because none of the modes dominates; the knowledge of both modes (representativity) is essential in order to conclude about the worst pWCET.

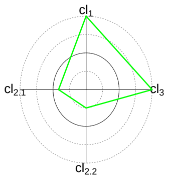

Figure 17 Radar plot for *trace6*.

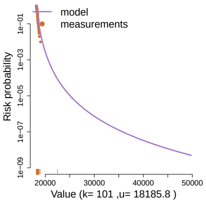

Figure 18 CCDF representation of *trace6* pWCET.

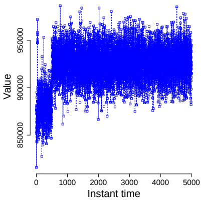

Figure 19 *trace7* with two operational modes.

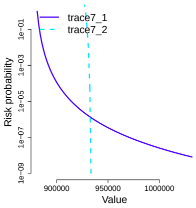

Figure 20 CCDF representation of *trace7* decomposed into *trace7_1* and *trace7_2*.

As a marginal note, we observe that DIAGXTRM infers two different shapes for the pWCET distributions of *trace7_1* and *trace7_2*; this is because DIAGXTRM applies a best fit procedure to the input measurements, and in [24] it has been demonstrated that only the best fit guarantees safe pWCET estimates. Only the best fit allows respecting the representativity of measurements and seeking for the best pWCET shape to cope with that.

When the EVT is not applicable to the full trace (trace-merging) i.e. *trace5* and *trace7*, the trace has to be decomposed into sub traces each characterizing a task mode; the representativity has to be preserved by not neglecting any sub trace/mode. Then, the EVT can applied to all the sub traces (envelope) and with the guarantee of having inferred the worst pWCET estimate.

5 Conclusions

MBPTA and EVT demand for representative trace of execution time measurements in order to provide safe and confident pWCET estimates.

In case of statistical dependence, changing execution time measurements and artificially create independence in order to have "better EVT applications" may cause the effect of reducing the safety of the pWCET estimates; this is not affordable with worst-case execution time estimates. The only allowed modifications to execution time traces are the conservative ones. Instead, with multi-mode tasks and multiple execution conditions, representative measurements have to include all the execution conditions in order to be able to infer the worst pWCET; execution conditions cannot be neglected, especially worst-case conditions.

Full coverage of system execution conditions and the dominance between execution conditions will be thoughtfully investigated in future work. The representativity of the measurements will be quantified including also the identical distribution hypothesis and other system parameters. Statistic metrics will be developed to identify pattern and execution behavior for real-time task. Moreover, the differences between BM EVT and PoT EVT will continue to be investigated. Special attention will be given to measurements robustness, measurements representativity and trace changes.

References

1 J. Abella, J. del Castillo, M. Padilla, and F. J. Cazorla. Extreme value theory in computer sciences: The case of embedded safety-critical systems. In *6th International Conference on Risk Analysis (ICRA)*, 2015.

2 Jaume Abella, Eduardo Quiñones, Franck Wartel, Tullio Vardanega, and Francisco J. Cazorla. Heart of gold: Making the improbable happen to increase confidence in MBPTA. In *26th Euromicro Conference on Real-Time Systems, (ECRTS)*, 2014.

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

4 Alan Burns and Stewart Edgar. Predicting computation time for advanced processor architectures. In *12th Euromicro Conference on Real-Time Systems (ECRTS)*, 2000.

5 Francisco J. Cazorla, Tullio Vardanega, Eduardo Quinones, and Jaume Abella. Upper-bounding program execution time with extreme value theory. In *WCET*, 2013.

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

7 Corentin Damman, Gregory Edison, Fabrice Guet, Eric Noulard, Luca Santinelli, and Jerome Hugues. Architectural performance analysis of FPGA synthesized LEON processors. In *Proceedings of the IEEE International Symposium on Rapid System Prototyping*, 2016.

8 J. L. Díaz, D. F. Garcia, K. Kim, C. G. Lee, L. L. Bello, J. M. López, and O. Mirabella. Stochastic analysis of periodic real-time systems. In *23rd of the IEEE Real-Time Systems Symposium (RTSS)*, 2002.

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

10 C. A. T. Ferro and J. Segers. Automatic Declustering of Extreme Values Via an Estimator for the Extremal Index. *Technical Report*, 2002.

11 M. K. Gardner. *Probabilistic analysis and scheduling of critical soft real-time systems*. PhD thesis, University of Illinois at Urbana-Champaign, Champaign, IL, USA, 1999. AAI9953022.

12 F. Guet, L. Santinelli, and J. Morio. On the reliability of the probabilistic worst-case execution time estimates. In *8th European Congress on Embedded Real Time Software and Systems (ERTS)*, 2016.

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

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

15 Stephen Law and Iain Bate. Achieving appropriate test coverage for reliable measurement-based timing analysis. In *28th Euromicro Conference on Real-Time Systems*, 2016.

16 M. R. Leadbetter. Extremes and local dependence in stationary sequences. *Stochastic Processes and their Applications*, 35, 1983.

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

18 George Lima, Dario Dias, and Edna Barros. Extreme value theory for estimating task execution time bounds: A careful look. In *28th Euromicro Conference on Real-Time Systems, (ECRTS)*, 2016.

19 C. L. Liu and J. W. Layland. Scheduling algorithms for multiprogramming in a hard-real-time environment. *Journal of the ACM*, 1973.

20 Cristian Maxim, Adriana Gogonel, Irina Asavoae, Mihail Asavoae, Liliana Cucu-Grosjean, and Walid Talaboulma. Reproducibility and representativity – mandatory properties for the compositionality of measurement-based WCET estimation approaches. In *The 9th International Workshop on Compositional Theory and Technology for Real-Time Embedded System (CRTS)*, 2016.

21 Vincent Nélis, Patrick Meumeu Yomsi, and Luís Miguel Pinho. The variability of application execution times on a multi-core platform. In *16th International Workshop on Worst-Case Execution Time Analysis, 2016*, pages 6:1–6:11, 2016.

22 R. Pellizzoni and M. Caccamo. Toward the predictable integration of real-time COTS based systems. In *Real-Time Systems Symposium. (RTSS). 28th IEEE International*, 2007.

23 L. Santinelli, J. Morio, G. Dufour, and D. Jacquemart. On the Sustainability of the Extreme Value Theory for WCET Estimation. In *14th International Workshop on Worst-Case Execution Time Analysis (WCET)*, pages 21–30, 2014.

24 Luca Santinelli, Fabrice Guet, and Jerome Morio. Revising measurement-based probabilistic timing analysis. In *2017 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*. IEEE, 2017.