Open Challenges for Probabilistic Measurement-Based Worst-Case Execution Time

Samuel Jiménez Gil, Iain Bate, George Lima, Luca Santinelli, Adriana Gogonel, and Liliana Cucu-Grosjean

***Abstract*—The worst-case execution time (WCET) is a critical parameter describing the largest value for the execution time of programs. Even though such a parameter is very hard to attain, it is essential as part of guaranteeing a real-time system meets its timing requirements. The complexity of modern hardware has increased the challenges of statically analyzing the WCET and reduced the reliability of purely measured the WCET. This has led to the emergence of probabilistic WCETs (pWCETs) analysis as a viable technique. The low probability of appearance of large execution times of a program has motivated the utilization of rare events theory like extreme value theory (EVT). As pWCET estimation based on EVT has matured as a discipline, a number of open challenges have become apparent when applying the existing approaches. This letter enumerates key challenges while establishing a state of the art of EVT-based pWCET estimation methods.**

AQ1 ***Index Terms*—**

I. Introduction

THE PROGRAMS of a real-time system should produce correct outputs computed within a time limit. To meet this constraint the worst-case execution time (WCET) of the running program is needed as an input to schedulability analysis. Unfortunately, determining the WCET of such a program is *intractable* as it would require knowledge of all possible *states* of the program [1]. Considering these constraints, the actual WCET is seldom known. Instead, what is achievable are WCET estimations based on assumptions of the system behavior. The WCET estimation methods should be *acceptably sound*, i.e., rarely optimistic without being overly pessimistic. In well designed systems the occasional underestimation can be tolerated as task deadlines would only be missed if other tasks also executed for times near their WCET and even if the deadlines are missed then the system has other levels of fault tolerance [2]. The number and pattern of allowable over estimations leads to a *target reliability* for WCET analysis. Too much pessimism means more budget has to be assigned to the task than needed which wastes system resources.

Manuscript received March 8, 2017; accepted May 21, 2017. This manuscript was recommended for publication by D. Sciuto. *(Corresponding author: Iain Bate.)*

S. Jiménez Gil and I. Bate are with the Department of Computer Science, University of York, York YO105GH, U.K. (e-mail: iain.bate@york.ac.uk).

AQ2 G. Lima is with the Department of Computer Science, Federal University of Bahia, Salvador, Brazil.

L. Santinelli is with the Department of Computer Science, ONERA, Mauzac, France.

A. Gogonel and L. Cucu-Grosjean are with INRIA, Le Chesnay, France.

Digital Object Identifier 10.1109/LES.2017.2712858

Classical WCET estimation techniques are based on *static timing analysis* which involves building an accurate model of both the underlying hardware and the program [2]. Modern hardware equipped with performance enhancement units have dramatically complicated the static modeling [3] leading to an interest in measurement-based techniques. As the larger values of execution time are often hard to create test cases for and in normal operation occur infrequently [4], the measurement-based approaches are combined with probabilistic models that quantify how likely an execution time is exceeded. As a result, a probabilistic WCET (pWCET) is obtained. These methods are known as measurement-based probabilistic timing analyses (MBPTA), whereas the *static probabilistic timing analysis* extends the static analysis to include probabilistic estimates. It is noted any measurement-based technique cannot by definition guarantee that the WCET is pessimistic or tight except in the simplest of cases.

The seminal work on estimating pWCET with an MBPTA approach is proposed by Burns and Edgar [5] and it is based on extreme value theory (EVT), a statistics branch advocated to the study of rare events. Despite several (and recent) developments on EVT-based MBPTA methods, important challenges exist. In this letter, we outline the state of the art for EVT-based MBPTA and the associated challenges. A short introduction to the EVT application to the estimation problem is given in Section II. A state of the art on EVT-based MPBTA methods is resumed in Section III followed by Section IV, where we identify the key research challenges ensuring the EVT applicability to the pWCET estimation problem.

II. Applying EVT to Execution Time Measures

Applying EVT to the pWCET estimation problem consists of different steps which are synthesized as follows.

1) Collecting the execution times from the system under test such that the identically distributed and/or independence hypotheses are satisfied for $(X_i)_1^n$, where $(X_i)_1^n$ is the set of measurements X_i, $i = 1, 2, \ldots, n$, obtained as the execution of a program.
2) Building a set of maxima from the set of execution times is done by selecting the maxima from $(X)_1^n$. Two classical methods of selection are block maxima (BM) and peaks-over-threshold (PoT). The former consists of partitioning the sampled data $(X)_1^n$ into equally sized blocks, whose sizes are specified beforehand, and selecting the maximum of each block; whereas the latter selects *all* values in $(X)_1^n$ above a certain previously

defined threshold. Both approaches involve the careful selection of a parameter, i.e., the block size or the threshold.

3) The EVT applicability is checked for the set of maxima by testing whether the sample of maxima converges to any one of the three possible extreme value (EV) distributions, e.g., Gumbel, Weibull, or Frechet under the BM approach.
4) Deriving an EV model is obtained by fitting the maxima set into either: a generalized EV distribution (GEV) when the set of maxima is selected using the BM principle; or a generalized Pareto distribution (GPD)when the set of maxima is selected using the PoT selection. In either case, their distribution parameters (e.g., shape, location, and scale) are obtained.
5) The validity of the model is checked in more recent papers by using some form of goodness-of-fitness test to check whether the obtained EV model describes the empirical sample of maxima. More recently Santinelli *et al.* [6] has defined a number of hypothesis to be checked as part of the steps as part of providing evidence that the result from the steps is valid.
6) Extracting a high quantile (i.e., probabilistic bound) from the obtained EV-model is done by determining a value $q(p)$ associated with a probability of exceedance, i.e., how likely the execution time is expected to be exceeded, p. That is, $\Pr\{X_i > q(p)\} = p$.

It is noted the probability of exceedance and related confidence intervals for the pWCET estimation derived via EVT is usually not the same as the likelihood the pWCET is exceeded in practice [7]. The reason is there are a number of uncertainties in the approach [8], e.g., the set of test cases will be incomplete, there are a number of parameters (e.g., the block size) which are tradeoffs, and the choice of distribution parameters is also a compromise.

III. State of the Art

In their seminal work [5], Edgar and Burns fit directly the top (i.e., the highest X%) of the execution times to the GEV distribution obtained as a combination of the three probability distributions defined as upper bounds by EVT. A key difference to the protocol in Section II is that neither BM or PoT is applied. A second work [9] from the same proposes the direct fitting of the top of the execution times to the Gumbel distribution. Edgar acknowledged later in his Ph.D. authors thesis [10] that a specific probability distribution, e.g., Gumbel, may not always be suitable for all programs.

In 2009, Hansen *et al.* [11] revisited the EVT application to the pWCET estimation problem. The quality of the Gumbel fitting method used is check by the χ^2-squared goodness-of-fit test. In 2012, Cucu-Grosjean *et al*. [12] and Wartel *et al.* [13] the next year, provide a detailed statistical analysis testing the Gumbel hypothesis using the "*Exponential Tail Test*" [12], [13]. This test replaces the χ^2 test as the latter was considered inadequate for distribution tail fitting. Indeed the χ^2 test focuses on the central part of the distribution while the interesting (pWCET) values are expected to be found in the tails.

The Gumbel and GEV hypotheses are enriched by using GPD distributions [14]–[16] indicating that the EVT application to the pWCET estimation problem is not restricted to the Gumbel and/or GEV distributions.

Independent of how the EVT approach is applied, the realism and applicability of EVT results is criticized by Griffin and Burns [17]. Their main concerns are the appropriateness of the input data and the validation of the results without a ground truth. To address this concern, Lesage *et al.* [18] developed a framework combining a proper set of hypothesis-driven experiments that provides a *ground truth* to be compared with the predicted pWCET. The framework assesses the quality of the EVT results (i.e., whether the pWCET upper bounds the WCET and with what pessimism) and the reliability of the EVT results (i.e., the quality of the EVT results needs to be consistently good and importantly poor quality results should be sufficiently rare). The framework also allows the user to understand the implications of imperfect conditions when applying EVT (e.g., the input sample to EVT is incomplete). This latter case is mainly due to incomplete test coverage either with respect to the structure of the program or to the quantity of test cases. To date, structural coverage has been used while testing the functional properties fulfilled by the programs and the most common criterion is branch coverage. Branch coverage is rarely sufficient alone and probabilistic approaches are proposed to complete such analysis in presence of EVT-based approaches. For randomized caches Kosmidis *et al.* [19] proposed the path upper bounding accounting for combinations of blocks that had not been executed during the measurement protocol. Ziccardi *et al.* [20] completed this approach through the Extended Path Coverage technique which targets full path coverage also for randomized caches.

Providing coverage relies also on a sufficient cardinal for the sample of execution times. For instance Cucu-Grosjean *et al.* [12] offered a first iterative method to determine such a cardinal without any proof of existence of such a cardinal. Moreover, any measurement-based approach may lead to uncertainties so Lu *et al.* [8] considered applying posterior statistical correction to the EVT application. Ostensibly Lu calculated the probability of exceedance used in EVT through a function of the target reliability for the WCET and the known uncertainties in the measurement and analysis protocol.

Finally, time-randomized architectures (TRA) [21] have been proposed to enable key assumptions (i.e., the measures in the sample are i.i.d) of EVT to be met. However, such architectures do not guarantee these assumptions are met nor solve the open problems defined in this letter.

IV. Challenges and Open Problems

The six stages outlined in Section II lead to the following three challenges if EVT analysis is to be successfully applied to the problem of pWCET analysis. In this section, these are considered in turn from which open problems are defined.

1) *Stage 1:* What is a representative input sample of execution times for EVT.

2) *Stages 2–5:* How can we ensure a trustable application of EVT for a representative input sample of execution times.
3) *Stage 6:* For a trustable application of EVT and on a representative input sample, how do we interpret the EVT result.

A. Representative Input Sample to EVT

The sample of execution times provided as input to EVT for a pWCET estimation is obtained using a measurement protocol. This measurement protocol describes the status of the program and of the processor for each measurement as well as their variations between different measurements. Ideally the resulting sample would be the same as the deployed system. This creates two problems. First, the longest paths in a piece of software deals with abnormal cases which would be dangerous to replicate in a real system (for example a car steering system dealing with a tyre blowout) and even hardware-in-the-loop testing is not entirely realistic. Second, even if some trials were performed on a real system then they would be limited so few extremal values might be obtained. Therefore, our definition of representative is that the sample contains cases similar to the deployed extremal situations and that these cases form a distribution that means EVT produces a pWCET that is acceptably sound. However, it is worth remembering two issues. First, the actual WCET is not generally known and so the soundness of the estimations may not be easily checkable. Second, the pWCET value *also* depends on the sample of observations supplied to the fitting method, the fitting method itself, the asymptotic properties of the resulting GEV or GPD distribution and the exceedance probability from which the pWCET is derived.

Based on the challenges in this section, we enumerate the following open problems.

- I1 How to determine the requirements for representativity in the context of EVT and the wider system.
- I2 How to generate test vectors to satisfy the need for representativeness.
- I3 How to identify the appropriate abstraction for the structure of the program and processor such that achieving sufficient coverage at the chosen abstraction gives a representative sample.
- I4 How to identify the common properties of programs and processors so that a sufficient cardinal for the sample can be justified.
- I5 How to identify incomplete representativity of the sample and assess its impact on the pWCET estimation.
- I6 How many execution times are needed in the sample for a given program, processor, and target reliability for the pWCET.

B. Trustable Application of EVT in Timing Analysis

Besides the problem of obtaining execution time samples and checking their representativeness mentioned in the previous section, some aspects related to applying EVT in time analysis may also impact the soundness of pWCET derivation. Santinelli *et al.* [22] showed how sensitive the pWCET is when selecting the maximal observations for the fitting process. Once the maximal observations are filtered EVT theory [23]–[25] dictates that these observations should belong to a continuous distribution and be i.i.d. However, in general there is no guarantee that a given sample of maxima can be described by an EV distribution even for i.i.d continuous data [26]. TRA-based randomization also aims to remove intrinsic data discreteness, ensuring or reducing independence and making more likely the applicability of EVT-based time analysis. However, there are scenarios, where EVT fails even if TRA-based randomized architectures are used [16]. As an alternative, randomization has recently been applied to data samples [27] so as to make samples EVT-compliant. This approach was shown to achieve the i.i.d. assumption more effectively than TRA for both standard benchmark software and real industrial case studies [4].

As for the fitting, well known and established estimation methods are based on the maximum likelihood estimator but it can only be applied when the shape parameter of the EV distribution obtaining during distribution fitting is above $-1/2$ [25]. Moment-based methods [28] are more general but computer-based procedures to estimate confidence intervals are needed [29]. Although, those topics are more related with EVT, not being specific to timing analysis, pWCET estimation is greatly sensitive to small variations of the method used. One reason for this is that usually one is interested in very small values of exceedance probability, mainly when it comes to critical systems. Recently, it has been observed that distinct implementations of the same fitting method may produce different pWCET estimations [30].

If it is assumed that the sample obtained may be not representative, it would be required that this lack of representativeness could be compensated. Speculatively speaking, a possible compensation biasing the fitting method toward the appropriate right-tail of EV distributions, however, this would be predicated on knowing what the distribution should be. To the best of our knowledge neither EVT nor MBPTA methods published to date offer systematic methods for accomplishing this kind of requirement.

For any method to be useful to industry, they must be reproducible. In the context of EVT, a method can be considered reproducible if for the same sample of execution times the same pWCET estimates is obtained. The reason for this requirement is in case of issues the reason behind a method's output must be understood which means it needs to be precisely recreated.

With respect to this second challenge we enumerate the following open problems.

- A1 How do we demonstrate that the methods to estimate EV model parameters (and their implementation) are sufficiently reliable.
- A2 How do we ensure that EVT application leads to a sound pWCET in the context of the available data and the requirements of the system.
- A3 How can we compensate for the lack of representativeness in the sample inorder to derive a sound pWCET.
- A4 How do we argue that such an application of EVT methods as part of pWCET analysis is reproducible.

C. Interpretation of the EVT Results

Assuming that we have considered the steps described so far the last issue is to actually select the pWCET from the tail of the distribution. The choice of value is a complex issue and not well understood problem [7]. There are a number of issues. On the requirements side, the value needs to be chosen such that the risk of system hazard events is acceptable. The complexity comes from the fact the likelihood of an individual pWCET being exceeded has to be considered in the context of all the other software tasks, the fault tolerance mechanisms designed into this part of the system, and all the other parts of the system that might contribute to the hazardous events. From a timing perspective, previous work [31], [32] has looked at understanding how often tasks meet their deadlines for a given profile of execution times. From a risk management perspective, the larger the extrapolation from the observations to the calculated pWCET the greater the level of uncertainty.

With respect to this third challenge we enumerate the following open problems.

- O1 How to understand the uncertainties within the overall measurement and analysis protocol.
- O2 How do we establish the exceedance probability to providing a sound WCET with manageable risks.
- O3 How do we schedule and develop a system in the presence of the derived pWCET.
- O4 How the process of deriving the pWCET affects the certification argument.
- O5 How to demonstrate an appropriate relationship between the pWCET estimate of a program and the timing behavior of the overall system.

V. Conclusion

This letter provides a review of the state of the art literature for deriving the pWCET of software using MBPTA with EVT methods. A number of open challenges have been identified that should be useful motivation for future research. It is noted that the set of challenges is not claimed to be complete.

References

[1] R. Wilhelm *et al.*, "The worst-case execution-time problem—Overview of methods and survey of tools," *ACM Trans. Embedded Comput. Syst.*, vol. 7, no. 3, p. 36, 2008.

[2] P. Graydon and I. Bate, "Realistic safety cases for the timing of systems," *Comput. J.*, vol. 57, no. 5, pp. 759–774, 2013.

[3] R. Kirner and P. Puschner, "Obstacles in worst-case execution time analysis," in *Proc. 11th IEEE Int. Symp. Object Orient. Real Time Distrib. Comput.*, Orlando, FL, USA, 2008, pp. 333–339.

[4] S. Law and I. Bate, "Achieving appropriate test coverage for reliable measurement-based timing analysis," in *Proc. Euromicro Conf. Real Time Syst.*, Toulouse, France, 2016, pp. 189–199.

[5] A. Burns and S. Edgar, "Predicting computation time for advanced processor architectures," in *Proc. 12th Euromicro Conf. Real Time Syst.*, Stockholm, Sweden, 2000, pp. 89–96.

[6] L. Santinelli, F. Guet, and J. Morio, "Revising measurement-based probabilistic timing analysis," in *Proc. IEEE Real Time Embedded Technol. Appl. Symp.*, 2017. AQ3

[7] D. Griffin, I. Bate, and B. Lesage, "Evaluating mixed criticality scheduling algorithms with realistic workloads," in *Proc. 3rd Int. Workshop Mixed Criticality Syst.*, 2015, pp. 24–29.

[8] Y. Lu, T. Nolte, I. Bate, and L. Cucu-Grosjean, "A statistical response-time analysis of real-time embedded systems," in *Proc. Real Time Syst. Symp.*, San Juan, Puerto Rico, 2012, pp. 351–362.

[9] S. Edgar and A. Burns, "Statistical analysis of WCET for scheduling," in *Proc. Real Time Syst. Symp.*, London, U.K., 2001, pp. 215–224.

[10] S. Edgar, "Estimation of worst-case execution time using statistical analysis," Ph.D. dissertation, Dept. Comput. Sci., Univ. at York, York, U.K., 2002.

[11] J. Hansen, S. A. Hissam, and G. A. Moreno, "Statistical-based WCET estimation and validation," in *Proc. 9th Int. Workshop Worst Case Execution Time (WCET) Anal.*, 2009, pp. 1–11.

[12] L. Cucu-Grosjean *et al.*, "Measurement-based probabilistic timing analysis for multi-path programs," in *Proc. 24th Euromicro Conf. Real Time Syst.*, Pisa, Italy, 2012, pp. 91–101.

[13] F. Wartel *et al.*, "Measurement-based probabilistic timing analysis: Lessons from an integrated-modular avionics case study," in *Proc. 8th IEEE Int. Symp. Ind. Embedded Syst.*, Porto, Portugal, 2013, pp. 241–248.

[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 *Proc. 19th Int. Conf. Embedded Real Time Comput. Syst. Appl.*, Taipei, Taiwan, 2013, pp. 22–31.

[15] F. Guet, L. Santinelli, and J. Morio, "On the reliability of the probabilistic worst-case execution time estimates," in *Proc. 8th Eur. Congr. Embedded Real Time Softw. Syst. (ERTS)*, 2016.

[16] G. Lima, D. Dias, and E. Barros, "Extreme value theory for estimating task execution time bounds: A careful look," in *Proc. Euromicro Conf. Real Time Syst.*, Toulouse, France, 2016, pp. 200–211.

[17] D. Griffin and A. Burns, *Realism in Statistical Analysis of Worst Case Execution Times* (OASIcs-OpenAccess Series in Informatics), vol. 15. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2010. AQ4

[18] B. Lesage, D. Griffin, F. Soboczenski, I. Bate, and R. I. Davis, "A framework for the evaluation of measurement-based timing analyses," in *Proc. 23rd Int. Conf. Real Time Netw. Syst.*, Lille, France, 2015, pp. 35–44.

[19] L. Kosmidis *et al.*, "PUB: Path upper-bounding for measurement-based probabilistic timing analysis," in *Proc. Euromicro Conf. Real Time Syst.*, Madrid, Spain, Jul. 2014, pp. 276–287.

[20] M. Ziccardi, E. Mezzetti, T. Vardanega, J. Abella, and F. J. Cazorla, "EPC: Extended path coverage for measurement-based probabilistic timing analysis," in *Proc. Real Time Syst. Symp. (RTSS)*, San Antonio, TX, USA, 2015, pp. 338–349.

[21] F. J. Cazorla *et al.*, "PROARTIS: Probabilistically analyzable real-time systems," *ACM Trans. Embed. Comput. Syst.*, vol. 12, pp. 1–26, May 2013.

[22] L. Santinelli, J. Morio, G. Dufour, and D. Jacquemart, "On the sustainability of the extreme value theory for WCET estimation," in *Proc. 14th Int. Workshop Worst Case Execution Time Anal.*, vol. 39. Ulm, Germany, 2014, pp. 21–30.

[23] E. J. Gumbel, *Statistics of Extremes*. Courier Corp., 2012.

[24] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events for Insurance and Finance* (Applications of Mathematics). Heidelberg, Germany: Springer, 1997. AQ5

[25] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, vol. 208. London, U.K.: Springer, 2001.

[26] D. Dietrich, L. Haan, and J. Hüsler, "Testing extreme value conditions," *Extremes*, vol. 5, no. 1, pp. 71–85, 2002.

[27] G. Lima and I. Bate, "Valid application of EVT in timing analysis by randomising execution time measurements," in *Proc. IEEE Real Time Embedded Technol. Appl. Symp.*, 2017.

[28] J. R. M. Hosking, "L-moments: Analysis and estimation of distributions using linear combinations of order statistics," *J. Roy. Stat. Soc.*, vol. 52, no. 1, pp. 105–124, 1990.

[29] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Boca Raton, FL, USA: Chapman & Hall, 1994.

[30] C. Maxim, A. Gogonel, I.-M. Asavoae, M. Asavoae, and L. Cucu-Grosjean, "Reproducibility and representativity—Mandatory properties for the compositionality of measurement-based WCET estimation approaches," in *Proc. 9th Int. Workshop Compositional Theory Technol. Real Time Embedded Syst.*, 2016, pp. 17–25.

[31] I. Bate, A. Burns, and R. I. Davis, "A bailout protocol for mixed criticality systems," in *Proc. Euromicro Conf. Real Time Syst.*, Lund, Sweden, 2015, pp. 259–268.

[32] I. Bate, A. Burns, and R. I. Davis, "An enhanced bailout protocol for mixed criticality embedded software," *IEEE Trans. Softw. Eng.*, vol. 43, no. 4, pp. 298–320, Apr. 2017.

AUTHOR QUERIES

AUTHOR PLEASE ANSWER ALL QUERIES

PLEASE NOTE: We cannot accept new source files as corrections for your paper. If possible, please annotate the PDF proof we have sent you with your corrections and upload it via the Author Gateway. Alternatively, you may send us your corrections in list format. You may also upload revised graphics via the Author Gateway.

AQ1: Please supply index terms/keywords for your paper. To download the IEEE Taxonomy, go to http://www.ieee.org/documents/taxonomy_v101.pdf.

AQ2: Please provide the postal code for "Federal University of Bahia, Salvador, Brazil," "ONERA, Mauzac, France," and "INRIA, Le Chesnay, France."

AQ3: Please provide the page range for References [6], [15], and [27].

AQ4: Please provide the publisher location for References [17] and [23].

AQ5: Please confirm if the location and publisher information for References [24] and [25] are correct as set.