Software Timing Analysis for Complex Hardware with Survivability and Risk Analysis

Sergi Vilardell[‡,†], Isabel Serra[†,⋆], Jaume Abella[†], Joan Del Castillo[Ψ], Francisco J. Cazorla [†]
[‡]Universitat Politecnica de Catalunya (UPC), Spain
[†]Barcelona Supercomputing Center (BSC), Spain
[⋆]Centre de Recerca Matematica (CRM), Universitat Autonoma de Barcelona (UAB), Spain
[Ψ]Departament de Matematiques, Universitat Autonoma de Barcelona (UAB), Spain

***Abstract*—The increasing automation of safety-critical real-time systems, such as those in cars and planes, leads to more complex and performance-demanding on-board software and the subsequent adoption of multicores and accelerators. This causes software's execution time dispersion to increase due to variable-latency resources such as caches, NoCs, advanced memory controllers and the like. Statistical analysis has been proposed to model the Worst-Case Execution Time (WCET) of software running such complex systems by providing reliable *probabilistic* WCET (pWCET) estimates. However, statistical models used so far, which are based on *risk* analysis, are overly pessimistic by construction. In this paper we prove that statistical *survivability* and risk analyses are equivalent in terms of tail analysis and, building upon survivability analysis theory, we show that Weibull tail models can be used to estimate pWCET distributions reliably and tightly. In particular, our methodology proves the correctness-by-construction of the approach, and our evaluation provides evidence about the tightness of the pWCET estimates obtained, which allow decreasing them reliably by 40% for a railway case study w.r.t. state-of-the-art exponential tails.**

***Index Terms*—probabilistic timing analysis, WCET, Weibull**

I. Introduction

The adoption of high-performance hardware is relentless to respond to unprecedented (guaranteed) performance needs of embedded software in automotive [1], space [2] and avionics [3]. This emanates from the increasing adoption of software-controlled autonomous systems in unmanned vehicles (e.g. autonomous cars, drones, and deep-space missions).

One of the most remarkable side effects of complex hardware/software platforms is that software execution times present high and unobvious dispersion due to variable-latency resources such as cache memories, networks-on-chip (NoCs), advanced memories and memory controllers, direct memory access (DMA) controllers, whose latency may vary due to intrinsic application behavior (e.g. data is allocated in different heap and stack locations across runs), interference across applications (e.g. contention in the NoC), and environmental conditions (e.g. contention with DRAM refreshes). For instance, Federal Aviation Administration (FAA) Report DOT/FAA/TC-16/51 [4] shows that the latency to load 4KB of data may vary by a factor of 3x on a moderately complex Freescale's quad-core P5040 processor.

The impact of execution time variability on Worst-Case Execution Time (WCET) estimates has been addressed using statistical analysis in the form of Measurement-Based Probabilistic Timing Analysis (MBPTA) methods [5], [6], [7], [8], [9], [10], [11]. MBPTA delivers an execution time distribution whose risk of exceedance upper-bounds the true risk of exceedance by construction. Hence, given an acceptable risk level for the particular application under analysis (e.g. related to the residual risk accepted in safety-critical systems [12]) expressed in the form of an exceedance probability P_{target}, MBPTA delivers the lowest execution time value whose exceedance probability $P_{estimate}$ is guaranteed to be exceeded at most with the target exceedance probability ($P_{estimate} \leq P_{target}$). For instance, the crosspoint in Figure 1 shows that the probability of a program to take more than 6,100 (cycles) is at most 10^{-3} under the Frechet distribution.

Extreme Value Theory (EVT) [13], appropriate for risk analysis, is used to model the right/upper tail of execution time distributions. In the context of probabilistic WCET (pWCET) estimation, *exponential tails* delivered by EVT, see Figure 1, have been shown by argument [7] and empirically [14] to provide a reliable tail model for pWCET estimation. However, tightness of those exponential tails is limited. *Light tails*, also shown in Figure 1, delivered by EVT (e.g. Reverse Weibull) have a maximum value that the tail approaches asymptotically and hence, those tails are theoretically the tightest ones. However, to our knowledge, no method has been devised so far to use them reliably for pWCET estimation. In this line, this paper addresses the limitations of EVT for pWCET estimation and proposes novel solutions provably reliable and delivering much tighter pWCET estimates than exponential tails. In particular, our contributions are as follows:

Analysis. We formally show that tail modelling in the context of *survivability analysis* [15] targets analogous questions to those of risk analysis. Then, we show that *log-concave*[1] distributions [16], inherited from survivability analysis, deliver

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO) under grant TIN2015-65316-P, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 772773), and the HiPEAC Network of Excellence. Jaume Abella has been partially supported by the MINECO under Ramon y Cajal postdoctoral fellowship number RYC-2013-14717.

[1]A function f is logarithmically concave log-concave for short, if the function $log(f(x))$ is concave.

 DOI 10.1109/ICCD46524.2019.00036

the tightest distribution models, but existing fitting methods fail to model tails [17], [18], needed for pWCET estimation.

Proposal. We propose the use of a subset of log-concave distributions: Weibull tail distributions with increasing hazard rate, i.e. with shape $\beta \geq 1$, neither Reverse Weibull as in EVT nor full Weibull as in survivability analysis. Our approach provides analogous accuracy to that of log-concave distributions, without limitations to extend them to arbitrarily low exceedance probabilities, as needed for pWCET estimation.

Assessment. We compare our method and EVT alternatives with a bootstrap analysis on large data samples (10^7 measurements) from a railway case study as ground truth. Our results provide evidence on the reliability of our approach and significant pWCET reductions w.r.t. exponential tails obtained with EVT, thus allowing to trustworthily increase system utilization noticeably.

II. Context and Extreme Value Theory

Critical systems undergo strict verification and validation (V&V) processes against their corresponding functional safety standards (i.e. ISO26262 in automotive [12] and DO178B/C[19] in avionics). Timing V&V processes require collecting evidence supporting that software will execute correctly and *timely*. In particular, those processes provide evidence that the risk of violating deadlines for critical real-time software is residual, since functional safety standards acknowledge that risk cannot be completely avoided. Thus, they impose thorough V&V processes that allow deeming such risk as "sufficiently low" so that it can be neglected. In this line, pWCET estimates allow to quantify such residual risk. Notably, statistical analysis is not new in domains like automotive. In fact, it is used during system analysis: for instance, random hardware failure rates and coverage are represented (and operated) with probabilities and percentages in the reference standard ISO26262 Part 5 [12]. On the software side, a probabilistic treatment of the residual risk of software faults has been shown to be compatible with ISO26262 [20].

EVT is the main framework to model high execution times used so far. EVT includes two distribution families. Generalized Extreme Value (GEV) distribution builds upon the convergence of block maxima (BM); and Generalized Pareto Distribution (GPD) builds upon peaks-over-threshold (PoT) [21], [13]. For the sake of illustration, we provide the Cumulative Distribution Function (cdf) for GPD:

$$F(x;\psi,\xi) = \begin{cases} 1 - \left(1 + \frac{\xi x}{\psi}\right)^{-1/\xi} & \xi \neq 0, \\ 1 - \exp\left(-\frac{x}{\psi}\right) & \xi = 0. \end{cases} \quad (1)$$

where ψ corresponds to the scale and ξ, the *extreme value index* (*evi*) in EVT, corresponds to the shape parameter. This parameter characterizes the maximum domain of attraction of the BM (which converges to the GEV family) and the PoT (which converges to the GPD family)[2]. In particular, $\xi > 0$ corresponds to heavy (Fréchet tails for GEV), $\xi = 0$ to exponential (Gumbel tails for GEV), and $\xi < 0$ to light (Reverse Weibull for GEV) tails. Figure 1 shows an example of the three tails for GEV.

[2]Note that a 3-parameter formulation for GPD exists, with an additional parameter (location). Such 3-parameter formulation is similar to that of GEV. In any case, the particular formulation used is irrelevant for our discussion.

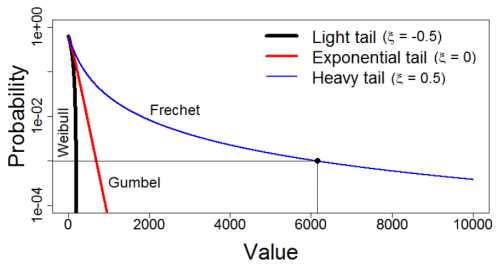


Fig. 1: Example of pWCET estimate obtained with light, exponential and heavy tail GEV distributions with $\xi = -0.5$, $\xi = 0$, and $\xi = 0.5$ respectively. ($\mu = 0$ and $\sigma = 100$). pWCET estimates are to be read as the maximum probability of the program to take longer than a given execution time value.

Authors in [22], [9] explore the use of unconstrained tail fitting, thus allowing ξ to take any value. As shown in [22], for heavy tails $\xi > 0$, pWCET bounds are inordinately large, becoming unusable in practice. Other authors rely on the fact that the execution time of real-time programs is finite to discard $\xi > 0$ by construction [23], [7], thus resorting to exponential (Gumbel) tails as the limit distribution that can be argued to be reliable, but much tighter than heavy (Fréchet) tails. Such assumption has been further verified empirically in [14]. We base our work on the premise that not only the execution time of real time programs is finite, but the probability of finishing a task increases as the task executes.

III. On the Use of Light Tails and Risk Analysis for WCET Estimation

A. Risk and Survivability Analysis

Tails lighter than exponential ones (so with $\xi < 0$) can deliver tighter bounds, as discussed in [7]. Yet, in the context of EVT, either GEV or GPD, distributions with $\xi < 0$ have a compact support, i.e. they have an absolute maximum value that cannot be exceeded. Hence, light tails in the case of EVT have an intrinsic risk of delivering optimistic tail distributions. This has some key implications in the fitting process, since a sufficiently large sample is needed to guarantee that light tail fitting is reliable for arbitrarily low exceedance probabilities. The target of our work is overcoming this limitation of the data and delivering a practical solution to obtain pWCET estimates tighter than those of exponential tails while preserving reliability. We do so by complementing EVT with survivability analysis as the theoretical ground for our hypothesis.

EVT is used in risk analysis to predict extreme (rare) events with the objective of proving that risk is below specific thresholds (e.g. financial risk). Survivability analysis, while it also focuses on predicting extreme events, it has opposite goals: proving that survivability is above specific thresholds (e.g. human life duration). Hence, both analyses target the modelling of extreme events, but with different objectives.

The type of distributions used to model those extreme events (i.e. tail distributions) differs across both analyses.

- EVT (either GPD or GEV) is used in the context of risk analysis, and it has been used so far in pWCET estimation

building on the idea that exceeding a specific execution time bound is a risk.
- For survivability analysis tail distributions can be split into two types: Decreasing Hazard Rate (DHR) and Increasing Hazard Rate (IHR) distributions. The boundary between those two categories corresponds to exponential distributions, which can be regarded as part of both.

In the context of pWCET estimation, we focus on IHR distributions, since they include those distributions that, as values get higher, the probability of realization increases. In our context this means that, as the program runs, there is an increasing probability of finishing the execution, which is the case of real-time programs that need to have a finite execution time to meet their deadline. Formally stated, a random variable X is IHR if the hazard rate function is increasing, where the hazard rate function is defined as:

$$h(x) = \frac{f(x)}{1 - F(x)}, \quad x \in support(X) \tag{2}$$

where f and F stand for the Probability Density Function (pdf) and the cdf of X, respectively. In Equation 2, $support(X)$ is a function representing the subset of the domain in which the random variable (X) probability is defined (i.e. it is not zero). In fact, the hazard rate function which assesses the IHR property is equivalent to the convexity H function, where $H(x) = -log(1 - F(x))$, called cumulative hazard rate.

B. IHR Distributions in Survivability Analysis

Log-concave distributions, as they can have an arbitrarily large number of parameters, are one of the best tools to fit data. On the other hand, they are generally defined only for the range of data observed. Hence, they are unable to model the distribution beyond that range, which is not useful for pWCET estimation. Some methods smooth the distribution by means of convolutions with other laws (e.g. Gaussian) to better fit the mode of the distribution [16], [17], [18]. While such an approach delivers a distribution that spans beyond the range of the data observed, it is tuned to model central behavior (not the tails) and so inherits the original problem we aim to tackle: an appropriate law needs to be identified for tail modelling.

Weibull distributions, among others, have often been used to model IHR distributions. However, they are unable to fit all tail distributions, so they are used only in specific contexts where the problem at hand matches the shape of those distributions [24], [16]. Therefore, while survivability analysis opens new opportunities to model pWCET, this is yet unexplored and existing distributions, in general, may not fit the needs of pWCET estimation. In this paper we address this challenge by contextualizing the needs of pWCET estimation and defining distribution families able to model pWCET distributions reliably, tightly and without incurring on the limitations imposed by log-concave distributions.

IV. Equivalence between IHR and Non-heavy Tails

In this section we introduce the connection between risk and survivability analysis in the context of pWCET estimation, which will lay out the ground for our hypothesis in further sections. pWCET estimates should be obtained theoretically under the assumption of the law of extreme events characterized by quicker decay in the tail than an exponential law, or equal, in the limit case. Exponential tails have been regarded as the appropriate (limit) model in practice [14], [7]. An exponential decay is a memoryless process where the probability of the process to complete is constant regardless of how long the process has been progressing. In the context of pWCET estimation, this corresponds to having a constant probability for the program to finish its execution despite the time elapsed since the program started running. Instead, the theoretical solution for pWCET estimation indicates that, the longer the program has been running, the higher the probability of finishing. That is, if X corresponds to execution time as a non-negative random variable, this assumption can be formally stated as follows if $s < t$ and $x > 0$:

$$P(X > t + x \mid X > t) \leq P(X > s + x \mid X > s), \tag{3}$$

In the context of extreme events, where u is the threshold upon which values belong to the tail of the distribution, s, t have to be large enough so that $s, t > u$, for some $u > 0$ [3]. In the general case of EVT (e.g. GPD), the tail decay can be described as follows for heavy, exponential and light tails respectively:

- if $\xi > 0$, then $P(X > t+x|X > t) \geq P(X > x)$ for $x > t$.
- if $\xi = 0$, then $P(X > t+x|X > t) = P(X > x)$ for $x > t$.
- if $\xi < 0$, then $P(X > t+x|X > t) \leq P(X > x)$ for $x > t$.

Note that the assumption described by Equation 3 for extreme events implies $\xi \leq 0$. Hence, the assumption for pWCET estimation matches the formulation above for light and exponential tails from GPD, inherited from risk analysis since, in the context of pWCET estimation, the longer the program has been running, the higher the probability of finishing. This matches the known concept in reliability modelling of IHR, which is therefore appropriate for pWCET estimation.

Building on Equation 3, and given a fixed x, we have the following equivalent assumption:

$$P(X < t + x|X > t) \geq P(X < s + x|X > s), \quad \text{if} \quad s < t$$

Given that $s < t$, then $P(X > t) < P(X > s)$, and hence, we can elaborate the equation above as follows:

$$\frac{P(X < t + x) - P(X < t)}{P(X > t)} \geq \frac{P(X < s + x) - P(X < s)}{P(X > s)}$$

If we use the cdf expressions instead, we have the excess distribution:

$$\frac{F(t + x) - F(t)}{1 - F(t)} \geq \frac{F(s + x) - F(s)}{1 - F(s)}, \quad \text{if} \quad s < t \tag{4}$$

If x tends to 0, then the cumulative probability ranges, between t and $t + x$ and between s and $s + x$, reduce to the particular probabilities at t and s respectively:

$$h(s) = \frac{f(s)}{1 - F(s)} \leq \frac{f(t)}{1 - F(t)} = h(t), \quad \text{if} \quad s < t \tag{5}$$

[3]Note that the threshold u is not larger than the theoretical threshold corresponding to the asymptotic behavior of the tail from the second fundamental theorem in EVT [25], [26]

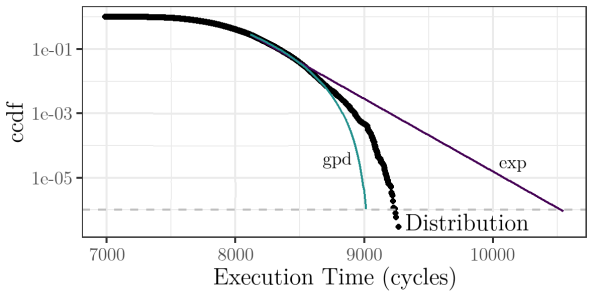


Fig. 2: Ccdf for GPD ($\xi < 0$) and exponential tails from EVT for TEST8. Both models are fitted with a sample of $n = 1000$ observations out of all $n = 10^7$ observations made.

As shown, Equation 5 – which we derive from Equation 3 – builds upon the hazard rate function shown in Equation 2. Hence, it models IHR distributions for survivability analysis, analogously to the GPD formulation for risk analysis. Note that in Equation 5 the equality case corresponds to a constant decay rate, hence a constant hazard rate function.

Log Concavity: In order to use IHR distributions for pWCET estimation, we build upon the following theorem proven in [15] and [27]:

Theorem. Given a non-negative random variable X, with f and F the pdf and cdf, respectively (where $H(x) = -\log(1 - F(x)), \quad x \in support(X)$),

$$\log(f) \text{ concave} \Rightarrow X \text{ IHR} \Leftrightarrow H \text{ convex} \tag{6}$$

Note that X is IHR in the tail, i.e. $(X \mid X > u)$ is IHR for some threshold $u > 0$, if and only if Equation 3 holds for all $s, t > u$ and, therefore, X is log-concave. Thus, by using log-concave distributions, IHR holds by construction.

A non-negative function is log-concave if its domain is a convex set, and if it satisfies the inequality $f(\theta x+(1-\theta)y) \geq f(x)^\theta f(y)^{1-\theta}$, for all x, y in the domain of f and $0 < \theta < 1$.

In order to test IHR one could make use of the log-concavity of the probability density function, which would give a convex H function and hence, IHR. Given an appropriate threshold u so that $(X \mid X > u)$ is IHR (and log-concave), we can fit a log-concave density function to the tail by using the maximum likelihood approach, as detailed in [17], [18]. Regardless of whether we fit the best log-concave distribution or a distribution function family preserving log-concavity but with much fewer parameters, as we do in this work, the exceedance threshold (u above) must be estimated to use the appropriate set of tail values from the sample for fitting. In particular, we build upon the work by Hazelton [27] that provides a procedure for testing whether we can reject the hypothesis of log-concavity for a given threshold u.

V. Weibull Tails (TailW) for pWCET Estimation

In Section II we have concluded that light tails with compact support are likely optimistic (thus unreliable) for pWCET estimation. On the other hand, exponential tails are the limit distribution for appropriate pWCET distribution models, hence being reliable but likely pessimistic. In order to further support this reasoning, Figure 2 shows an example of one task belonging to the railway case study we use in this paper, see Section VII. We fit GPD with the best fit, which naturally is a light tail (so $\xi < 0$), and an exponential (*exp*) distribution, using a sample with 1000 observations. In the figure, we depict the pWCET distributions obtained with light (GPD) and exponential tails, in the form of a complementary cdf (ccdf), as well as the empirical ccdf of a much larger data sample with 10^7 observations, which we use as ground truth. The cutoff values at an exceedance probability of 10^{-6} per run are also marked. As shown, the decay of the data sample sharpens for higher execution times, thus reflecting an IHR. The GPD best fit due to its compact support becomes eventually optimistic since the pWCET distribution reaches higher execution times for decreasing exceedance probabilities. The exponential tail, instead, has a fixed decay rate, hence getting farther away from the actual distribution as the exceedance probability decreases. Overall, GPD (with $\xi < 0$) and *exp* ($\xi = 0$) from EVT produce lower and upper bounds to the pWCET distribution for decreasing exceedance probabilities. Hence, we need an alternative model that has to satisfy the following properties:

1) Must have IHR in the tail (so H convex in the tail), thus having positive memory and $evi < 0$.
2) Must not have bounded (compact) support, not to suffer the same problems as GPD (with $evi < 0$).

The set of H-convex probabilistic models, i.e. with log-concave densities, satisfies the properties above and includes *all* probabilistic models satisfying those properties. However, as explained in Section III, those distributions may have a large number of parameters, which reduces the number of degrees of freedom. Furthermore, the fitting process allows describing them only for the probability range where data exists, which is useless for pWCET estimation.

Weibull distributions[4] have been often used to model survival processes such as, for instance, the lifetime of processors [24]. Failure rates over processor lifetime are usually shown in the form of a bathtub curve, where the failure rate decreases during the beginning of the lifetime (so DHR), since failures due to infant mortality are frequent. However, the more the processor survives in this phase, the lower the hazard rate. Eventually, a near-flat phase is reached where the hazard rate is nearly-constant, until the end of life period is approached, when the hazard rate increases (so IHR) until the processor eventually fails. Obviously, such a distribution does not meet the properties indicated above since it should be IHR, and Weibull distributions may have DHR for at least part of their support. However, if $\beta > 1$, where β stands for the shape, Weibull tails (tailW) are IHR and allow covering all the spectrum between GPD with $\xi < 0$ and exponential tails. In fact, if we allow $\beta \geq 1$, the boundary case where $\beta = 1$ corresponds to the exponential tails.

A. Formal Definition of tailW

The *tailW* law is constructed using the excess probability function, shown in Equation 4. Thus, the cdf is:

$$F(x, \alpha, \beta, \nu) = 1 - \exp\big(-\alpha(x + \nu)^\beta + \alpha\nu^\beta\big) \tag{7}$$

[4]The Weibull law for X is given by the cdf $F(x, \psi, \beta) = 1 - \exp\big(-(x/\psi)^\beta\big)$

for $x \geq 0$, $\alpha > 0$, $\beta \geq 0$ and $\nu > 0$. We consider *tailW* law with ν fixed and $\beta \geq 1$. The former reduces the cost of parameter estimation (only 2 parameters need to be estimated instead of 3) at the expense of delivering negligibly more pessimistic tail models. The latter ($\beta \geq 1$), as explained before, restricts *tailW* distributions to the domain of IHR. The likelihood ratio in the tail is described by:

$$l(x;\alpha,\beta,\nu) = n(\log(\alpha)+\log(\beta)) + (\beta-1)\sum_{i=1}^{n}\log(x+\nu) - \alpha\sum_{i=1}^{n}\left((x+\nu)^{\beta}-\nu^{\beta}\right)$$

and the MLE to fit the *tailW* law to the tail is obtained with numerical methods. The full definition of the *tailW* and a numerical method based on MLE to estimate its parameters can be found in the R package *distTails* [28], [29].

Since the purpose of tail prediction is only modelling tails, for which we lack sufficient empirical data to rely on the empirical quantile, we need to fit an appropriate law for the tail. However, for the rest of the distribution we can simply rely on empirical data. Hence, we can resort to a semi-parametric model, where for a fixed threshold u, the law for $x < u$ is given by the empirical law and for $x > u$ the law is given by a parametric model (e.g. *tailW* or *exp*).

VI. Fitting protocol

In order to use *tailW* distributions, we define an application protocol that guarantees reliability and maximizes tightness. Consider a sample, $\boldsymbol{x}$, $\{x_1,\ldots,x_n\}$ and a fixed threshold u to define the tail. We start by checking that the sample preserves the IHR property (log-concavity) for the considered exceedance threshold u as described in [27]:

1) Apply the bootstrap log-concavity test in the tail of the sample for the given u.
2) If the null hypothesis of log-concavity is rejected with risk 0.05, take another sample and restart.

Note that the significance level is 0.05, so that 95% of the bootstrapped samples must pass the test, is a common value for statistical tests. From this point onwards, log-concavity in the tail is assumed since it has already been tested. Also, $u \geq u_{EVT}$, where u_{EVT} is the threshold such that the theoretical approach from EVT can be applied. Then, the protocol continues as follows:

3) Fix $\lambda = F_{emp}(u)$
4) Consider the sample $\boldsymbol{y}$ given by $y_i = x_i/u - 1$ for i such that $x_i \geq u$.
5) Fit the *exp*, *tailW* and *logc* by MLE. Then, the loglikelihood for each law can be denoted by (where $\widehat{\Theta}$ corresponds to the set of parameters for a *logc* distribution)

$$\widehat{l}_{exp} = l_{exp}(\widehat{\psi};\boldsymbol{y}) \quad \widehat{l}_{tailW} = l_{tailW}(\widehat{\alpha},\widehat{\beta};\boldsymbol{y},\nu=1)$$
$$\widehat{l}_{logc} = l_{logc}(\widehat{\Theta};\boldsymbol{y})$$

6) Test the null hypothesis $tailW \sim exp$ through the Likelihood Ratio Test (LRT) [30], [31] with risk $\alpha = 0.05$:

$$2(\widehat{l}_{tailW} - \widehat{l}_{exp}) < \chi^{2,1}_{0.95}$$

Note that, since *tailW* has 2 parameters and *exp* 1, the χ^2 test is applied with 1 degree of freedom (the difference). If it is true, then the *exp* model (PoT with *exp* in the tail) must be considered for high-quantile estimation, and the fitting process finishes since the simplest model must be used if the models are not proven to differ[5]. Else, we continue with the next step.

7) Test the null hypothesis $logc \sim tailW$ with risk 0.05:

$$2(\widehat{l}_{logc} - \widehat{l}_{tailW}) < \chi^{2,\delta}_{0.95}$$

where δ is the number of parameters in *logc* fit minus 2 (those for *tailW*). If it is true, then *tailW* model (PoT with *tailW* law in the tail) must be considered for high-quantile estimation, and the fitting process finishes. Else, *tailW* may not be a sufficiently good fit. Hence, either we continue searching for *tailW* fitting with larger samples or another valid value for u, or we resort to the *exp* model (computing u_{EVT} and fitting the new tail as indicated before), which is known to be pessimistic but reliable for sky-high quantiles[6].

Overall, this application protocol is reliable by construction and aims at maximizing pWCET tightness.

VII. Evaluation

While theoretically *tailW* distributions meet the properties needed to model pWCET distributions tightly and reliably, we verify empirically such hypotheses in several ways:

- We compare *tailW* distributions against sky-high quantiles for large (ground truth) data samples.
- We compare *tailW* against GPD (with $\xi < 0$) and *exp*.
- We apply LRT to compare *tailW* and the reference *logc*.

Case study. We evaluate *tailW* with a railway case study running on an FPGA prototype. The railway case study is a safety-related function from the European Train Control System (ETCS) reference architecture in charge of distance supervision and travelling speed control. This function has strict real-time requirements and needs to be certified at the highest integrity level in the corresponding railway safety standards. We build upon 10 input data vectors, regarded as representative by the end user, which trigger different combinations of speed and acceleration, among other parameters. For each input, which we name from TEST0 to TEST9, we collect an execution time sample. In particular 10^3 measurements showed to be enough for each test. However, this would allow us deliver pWCET estimates but we would not have specific references to assess their reliability and tightness other than those predictions obtained with exponential tails. Thus, for the sake of having some form of ground truth, we have collected a large execution time sample with 10^7 measurements per TEST since they allow us assessing the quality of the pWCET estimates at sky-high quantiles. It is also worth mentioning

[5]If an *exp* tail is used for modelling, then EVT should be used to improve the fit, computing u_{EVT} such that $u_{EVT} \leq u$, and fitting the new tail.

[6]Note that typically, high quantiles are defined at the range 0.9-0.9999. Since we target very small exceedance probabilities, in line with safety standards requirements, we define sky-high quantiles those reaching values around $1-10^n$, where typically $n \in [-6,-12]$.

that collecting those data samples required 2 non-stop weeks of data collection, thus far beyond the typical effort an end user would spend for the timing analysis of a single program.

Platform. The platform for this experiment is an FPGA implementation of a LEON3+ architecture comprising first level instruction and data caches, and a unified L2 cache, where the sources of execution time variation have been conveniently controlled by hardware means to guarantee the representativeness of the measurements collected at analysis w.r.t. the timing behavior during operation [32], [33]. In particular, this processor builds upon the concepts of time upper-bounding and time randomization to enforce representativeness. However, our analysis is agnostic of the platform on which measurements are collected.

A. Assessing Model Hypotheses: H-Convexity and Light Tails

tailW relies on the convexity of the H-plot and lightness of the tail for the distribution modelled. By definition of our problem at hand (finite execution times), both properties must hold. Figure 3 shows the H-plot for the full samples of the railway case study (10^7 measurements). As shown graphically in the plots, all data sets are H-convex, thus in line with the hypothesis in Equation 3. The lightness of the tail can be assessed with the Coefficient of Variation (CV) statistic. Given that the CV in the context of the *gpd* for a certain threshold u is $\text{CV}(X_u) = 1/\sqrt{1-2\xi}$, where ξ is the *evi*, we are able to classify the nature of the tail. This method is properly developed and explained in the context of MBPTA in [7]. In Figure 4, we show the CV plot for all TEST traces, where the thick line corresponds to the trace for TEST6. The CV plot shows that, given that $\text{CV} \leq 1$ in general, distributions have light tails. Small discontinuities in the data sample, which may happen due to random sampling, may create peaks when a low number of exceedances is considered; as in the case of TEST0 and TEST 4, in Figure 4. Regarding TEST6, it is the only one with a slightly different behavior since the leftmost part of its CV is slightly heavy ($\text{CV} \geq 1$), and it only becomes light after excluding half of the sample. We discuss TEST6 in more detail later.

B. Assessment with Large Data Sets

To assess the reliability and tightness of the *tailW* model, for each of the 10 TESTs, we conduct a bootstrap experiment consisting of generating 1000 random samples with 1000 observations each from the 10^7 observations collected for that TEST. Then, we fit the exponential (*exp*), the GPD with $\xi < 0$ and *tailW* models for each of those 1000 data samples. The pWCET value obtained for each of the three methods is assessed against the empirical distribution sky-high quantile $1 - 10^{-6}$ (so at an exceedance probability of 10^{-6}). Note that, by building on a data sample with 10^7 measurements, the highest quantile we could consider would be $1 - 10^{-7}$, which would be fully dependent on the highest value in the sample. Relying on a single (randomly sampled) value may bring some instability, so we opted for considering a lower (still sky-high) quantile at $1 - 10^{-6}$.

Figure 5a shows the Quantiles Of BootStrap (QOBS) estimator for TEST0 in the form of a boxplot. All other

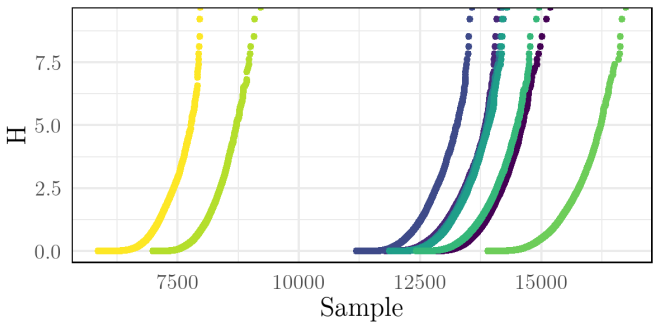


Fig. 3: H-plots for the whole TEST0 and TEST6 for a bootstrap sample of 10000 observations. The x-axis shows execution times in processor cycles.

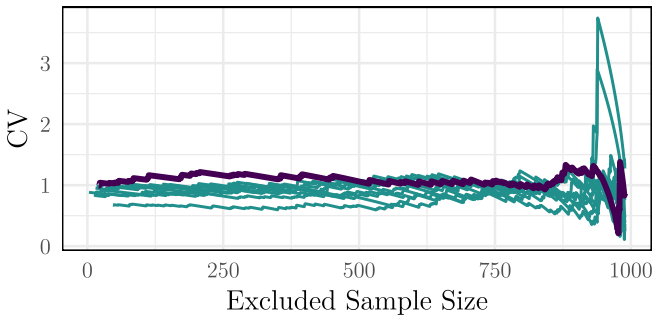
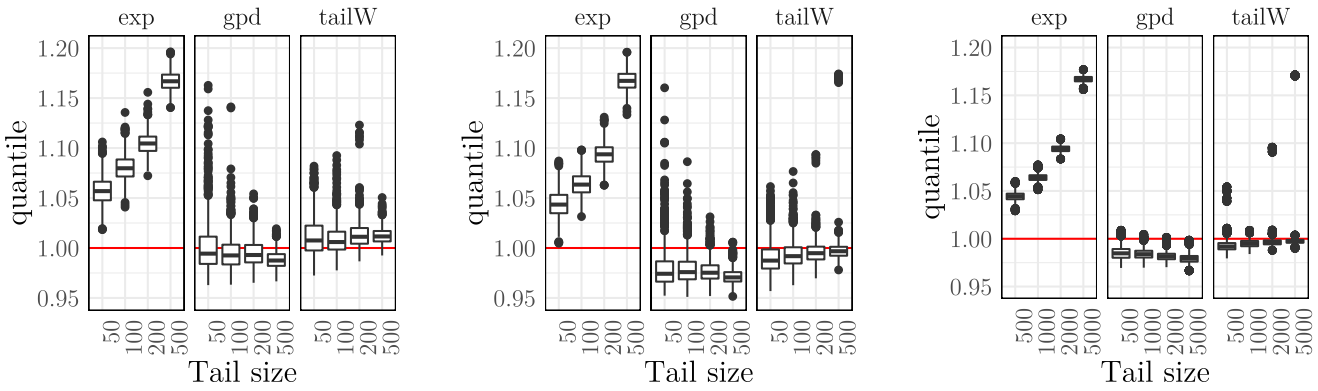


Fig. 4: Plot of the CV against the excluded sample size. The thick purple line corresponds to the TEST6, while the thin lines come from the rest of the tests. The CV plot was computed using the R package *ercv* [34].

TESTs except TEST6 have analogous behavior to TEST0, so we omit them for space needs. Results are obtained using different number of extremes (maxima) for tail fitting: we have used 500, 200, 100 and 50 extremes selected with PoT. The assumption formalized in Equation 3 lets us use the hypothesis on IHR for all data. Therefore, those numbers of extremes (between 50 and 500) can be reliably used for our analysis. As shown, *exp* provides highly pessimistic estimates with 500 extremes, whose mean is in the range $[1.15, 1.20]$. Tightness improves as we decrease the number of extremes. However, fitting a distribution with a lower number of measurements brings increased uncertainty.

As expected, *gpd* tends to underestimate the sky-high quantile of the real data with a low number of extremes (50), and quantiles are wide. Hence, in this case uncertainty is relatively high and confidence intervals should be wide. By increasing the number of extremes, uncertainty rapidly decreases and quantiles narrow down. However, *gpd* further underestimates the sky-high quantile of the real data due to losing the asymptotic tail behavior by using values farther away from the maximum in the sample.

Finally, *tailW* tends to tightly and reliably estimate the sky-high quantile of the real data with few extremes (50), but with wide quantiles. As we increase the number of extremes up to 500, we observe that reliability and tightness are preserved,



(a) TEST0 1000 sample (b) TEST6 1000 sample (c) TEST6 10000

Fig. 5: QOBS estimator distribution for TEST0 and TEST6 for 1000 and 10000 samples, under different models: *exp*, *gpd* and *tailW*, with different number of extremes: (500, 200, 100, 50) for samples of 1000 measurements and (5000, 2000, 1000, 500) for 10000 measurements.

TABLE I: Lower confidence interval (LCI) for the reference values for all railway TESTs.

Test	LCI	Test	LCI
TEST0	0.00%	TEST5	0.28%
TEST1	0.08%	TEST6	0.97%
TEST2	0.13%	TEST7	0.10%
TEST3	0.17%	TEST8	0.21%
TEST4	0.00%	TEST9	0.35%

and quantiles quickly narrow down. Overall, *tailW* provides much tighter (and still reliable) pWCET estimates than *exp*, and does not suffer the underestimation problems of *gpd* due to the compact support of *gpd*, which should naturally worsen as we consider higher sky-high quantiles (a.k.a. lower exceedance probabilities). Note that, sporadically, *tailW* may lead to pessimistic sky-high quantile estimates (comparable to those for *exp*). As explained before, occasionally, *tailW* fitting may not be sufficiently good and then, our model resorts to exponential tail fitting to preserve reliability.

For TEST6 (Figure 5b), the larger the number of extremes considered (and so the higher the confidence), the closer the estimator to the reference value for *tailW*, but still most of the distribution is below the reference value. For larger samples of 10000 measurements – Figure 5c – estimates for TEST6 become more precise, but still slightly below the reference value (up to 1%). Since the reference value is obtained with point estimation, we have calculated its lower confidence interval (a.k.a. how much reference lines at 1.0 should be moved down in the y-axis). In particular, we use the *binomial confidence interval* [35], since it only requires the size of the sample used and the number of successes/failures. We compute the 95% confidence interval. Results for all TESTs, see Table I, show that all confidence intervals are tiny except for TEST6, whose lower confidence interval is $\approx$1%, which means that *tailW* estimates are within the confidence interval of the ground truth value.

Since standards accept failure rates in the order of 10^{-5} to 10^{-9} failures per hour, and critical real-time tasks may run up to several thousands of times per hour, we consider exceedance probabilities of 10^{-6} and 10^{-12} per run for the pWCET estimates, thus showing the sensitivity of the different methods to the exceedance probability. In particular, we compare the only two reliable methods: *exp* and *tailW*. Given that *gpd* has been proven unreliable, we do not consider it for pWCET estimation. Therefore, we estimate the pWCET at such probabilities for *tailW* and *exp* with samples of 1000 execution time measurements. We consider, as in previous experiments, different numbers of extremes (50, 100, 200 and 500), and show the results normalized w.r.t. *tailW*.

As shown in Figure 6, *exp* delivers pWCET estimates between 5% and 20% higher than those of the proposed *tailW* method for the railway case study for an exceedance probability of 10^{-6} per run. As shown in Figure 5, tighter estimates are obtained for *exp* if fewer extremes are considered, whereas *tailW* accuracy is highly insensitive to this parameter. However, uncertainty increases if the number of extremes used is relatively low. Therefore, as the reliability required for the pWCET estimate increases, *tailW* provides higher gains.

Results for an exceedance probability of 10^{-12} per run are shown in Figure 7. Such a lower probability should be used for more critical tasks and/or for those tasks running more often. As shown, trends are similar to those of 10^{-6} exceedance probability, but at a higher scale, since *exp* pWCET estimates are typically between 15% and 50% higher than those of *tailW*. This increasing gap between the (tight) *tailW* model and *exp* model can be easily understood looking at Figure 2, where we see that the gap between *exp* and the actual distribution increases at decreasing exceedance probabilities. Note that achieving higher savings for tasks running more frequently implies that potential savings in system utilization can also be larger.

C. *Comparing* exp, tailW *and* logc *Models*

As explained before, *logc* distributions are the reference model, but they can only be used in the value range determined by input data. Nevertheless, we assess whether *tailW* delivers

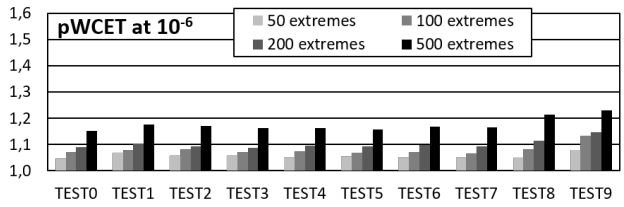


Fig. 6: pWCET estimate increase for *exp* w.r.t. *tailW* at an exceedance probability of 10^{-6} per run for the railway case study with different numbers of extremes for the tail.

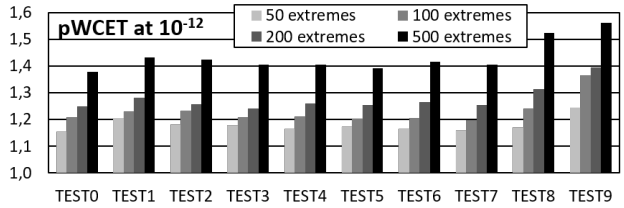


Fig. 7: pWCET estimate increase for *exp* w.r.t. *tailW* at an exceedance probability of 10^{-12} per run for the railway case study with different numbers of extremes for the tail.

distributions that cannot be distinguished from *logc* ones statistically in the range where the latter are defined. For that purpose, we perform an LRT.

We have applied the LRT to the 10 TESTs on the large data sets. As shown in Table II, the test is passed in almost all cases. In particular, the p-value is below 0.05 only for TEST1 (T1) with 100 extremes and TEST4 (T4) with 500 extremes. Hence, this result confirms that *M1* (so *tailW*) cannot be distinguished from *logc* distributions, thus supporting the high accuracy of the proposed model. In fact, in those cases where the test is failed, we only need to select an appropriate number of extremes that ensures that *tailW* is indistinguishable. The default solution consists on collecting a new sample since both distributions are naturally indistinguishable. Note that, in general, α determines the ratio of false negatives (a.k.a. wrong hypothesis rejections). In our case, given that $\alpha = 0.05$, we would expect 2 rejections out of 40 tests, which is exactly what we obtained.

For completeness, we have applied the LRT to compare *exp* with *tailW*. A test pass would mean that the simpler model (*exp* has just 1 parameter whereas *tailW* has 2) should be used instead of the complex one. Our results (omitted due to space constraints) show that the test is failed in most of the cases, thus meaning that *exp* is unable to capture tail distributions with as much accuracy as *tailW*.

TABLE II: LRT p-values for the 10 TESTs comparing *tailW* and *logc* models.

	tailW vs *logc*			
	500	200	100	50
T0	0.91	0.63	0.24	0.19
T1	0.07	0.08	0.04	0.72
T2	0.72	0.63	0.21	0.37
T3	0.21	0.57	0.30	0.94
T4	0.01	0.72	0.48	0.80
T5	0.98	0.25	0.67	0.19
T6	0.88	0.77	0.97	0.97
T7	0.75	0.85	0.34	0.59
T8	0.52	0.61	0.25	0.51
T9	0.11	0.88	0.21	0.92

VIII. Related Work

EVT relies on i.i.d. observations in the input sample [36], [37], [38]. The difficulty of achieving this depends on the underlying hardware platform [39], [33], [40], [10] and the software support used [8]. For instance, the COTS hardware platforms considered in [40], [10] pose difficulties to enforce i.i.d. measurements, and some dependencies across measurements may exist, thus requiring special consideration, as shown by Santinelli et al. [10].

The source of dependencies has been investigated by Melani et al. [41] concluding that pWCET estimation is still possible despite dependencies. Other authors build upon alternative methods for measurement collection to get rid of dependencies. In particular, Yue et al. [40] consider retaining only maxima for that purpose. Lima and Bate [8] show that mitigating the impact of discrete data may also help to mitigate the impact of dependencies.

Several authors have considered general EVT distributions, without restricting them to the exponential law [9], [8], [22], whereas others build on the particular characteristics of the problem modelled – finiteness of the execution time of critical real-time programs – to fit exponential tails only [42], [37], [23], [7], which have been shown to provide usable bounds and confidence intervals [14].

MBPTA brings several challenges [43] including a sound tail modelling application – the target of this work, the representativeness of the input samples [11], [44], [39], [10], [22], and how to interpret the obtained results. These three challenges have been presented in [43], while [45] makes a deep survey of the existing works addressing each of these challenges. In this context, [20], [46] have shown how to interpret EVT results from a certification point of view.

IX. Conclusions

While light tails are the theoretical best fit for pWCET estimation, the lack of reliable fitting methods (for light tails as part of EVT) that deliver reliable pWCET estimates for arbitrarily low exceedance probabilities imposes the use of exponential tails, which although proven reliable, are increasingly pessimistic for decreasing exceedance probabilities.

We proved that risk analysis, where EVT is used, and survivability analysis tackle the same fundamental problem by predicting the occurrence of rare events. Then, building on distributions from survivability analysis, we propose the use of Weibull tails (*tailW*), which are proven to be as reliable as reference log-concave distributions, but enabling the modelling of arbitrarily low exceedance probabilities.

We validate our approach based on *tailW* on a railway case study, showing that *tailW* provides reliable and tight pWCET estimates, as opposed to EVT light tails (often unreliable) and exponential tails (often pessimistic). Our results show that pWCET estimates can be reduced by around 40% w.r.t. current practice based on exponential tails, thus enabling much higher utilization of hardware platforms.

REFERENCES

[1] ARM, "ARM Expects Vehicle Compute Performance to Increase 100x in Next Decade," https://www.arm.com/about/newsroom/arm-expects-vehicle-compute-performance-to-increase-100x-in-next-decade.php, 2015.

[2] A. West, "NASA Study on Flight Software Complexity. Final Report," NASA, Tech. Rep., 2009.

[3] M. P. et al, "Mixed-criticality embedded systems - A balance ensuring partitioning and performance," in *Euromicro DSD*, 2015.

[4] William J. Hughes, "DOT/FAA/TC-16/51. Assurance of Multicore Processors in Airborne Systems." Federal Aviation Administration, Tech. Rep., 2017.

[5] F. Cros, L. Kosmidis, F. Wartel, D. Morales, J. Abella, I. Broster, and F. J. Cazorla, "Dynamic software randomisation: Lessons learnec from an aerospace case study," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*, March 2017, pp. 103–108.

[6] M. Fernandez, D. Morales, L. Kosmidis, A. Bardizbanyan, I. Broster, C. Hernandez, E. Quinones, J. Abella, F. Cazorla, P. Machado, and L. Fossati, "Probabilistic timing analysis on time-randomized platforms for the space domain," in *Design, Automation Test in Europe Conference Exhibition (DATE), 2017*, March 2017, pp. 738–739.

[7] 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, pp. 72:1–72:29, Jun. 2017. [Online]. Available: http://doi.acm.org/10.1145/3065924

[8] G. Lima and I. Bate, "Valid application of evt in timing analysis by randomising execution time measurements," in *2017 IEEE Real-Time and Embedded Technology and Applications Symposium (RTAS)*, April 2017, pp. 187–198.

[9] G. Lima, D. Dias, and E. Barros, "Extreme value theory for estimating task execution time bounds: A careful look," in *2016 28th Euromicro Conference on Real-Time Systems (ECRTS)*, July 2016, pp. 200–211.

[10] 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*, ser. OpenAccess Series in Informatics (OASIcs), H. Falk, Ed., vol. 39. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2014, pp. 21–30. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2014/4601

[11] J. Abella, C. Hernandez, E. Quiñones, F. J. Cazorla, P. R. Conmy, M. Azkarate-askasua, J. Perez, E. Mezzetti, and T. Vardanega, "Wcet analysis methods: Pitfalls and challenges on their trustworthiness," in *10th IEEE International Symposium on Industrial Embedded Systems (SIES)*, June 2015, pp. 1–10.

[12] International Organization for Standardization, *ISO/DIS 26262. Road Vehicles – Functional Safety*, 2009.

[13] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*. Springer, 2001.

[14] K. P. Silva, L. F. Arcaro, and R. S. d. 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.

[15] D. Cox and D. Oakes, *Analysis of Survival Data*, ser. Monographs on Statistics and Applied Probability. Chapman and Hall, 1984.

[16] A. Marshall and I. Olkin, *Life distributions. Structure of Nonparametric, Semiparametric, and Parametric Families*, ser. Springer Series in Statistics. Springer, 2007.

[17] F. Balabdaoui, K. Rufibach, and J. A. Wellner, "Limit distribution theory for maximum likelihood estimation of a log-concave density," *Ann. Statist.*, vol. 37, no. 3, pp. 1299–1331, 06 2009. [Online]. Available: https://doi.org/10.1214/08-AOS609

[18] L. Duembgen, A. Huesler, and K. Rufibach, "Active set and EM algorithms for log-concave densities based on complete and censored data," IMSV, Univ. of Bern, Tech. Rep., 2010. [Online]. Available: http://arxiv.org/abs/0707.4643v4

[19] RTCA and EUROCAE, *DO-178C / ED-12C, Software Considerations in Airborne Systems and Equipment Certification*, 2011.

[20] I. Agirre, F. J. Cazorla, J. Abella, C. Hernández, E. Mezzetti, M. Azkarate-askatsua, and T. Vardanega, "Fitting software execution-time exceedance into a residual random fault in ISO-26262," *IEEE Trans. Reliability*, vol. 67, no. 3, pp. 1314–1327, 2018. [Online]. Available: https://doi.org/10.1109/TR.2018.2828222

[21] S. Kotz and S. Nadarajah, *Extreme value distributions: theory and applications*. World Scientific, 2000.

[22] F. Guet, L. Santinelli, and J. Morio, "On the Reliability of the Probabilistic Worst-Case Execution Time Estimates," in *Embedded Real-time Software and Systems (ERTS^2) Conference*, 2016.

[23] L. Cucu-Grosjean, L. Santinelli, M. Houston, C. Lo, T. Vardanega, L. Kosmidis, J. Abella, E. Mezzetti, E. Quiñones, and F. J. Cazorla, "Measurement-based probabilistic timing analysis for multi-path programs," in *2012 24th Euromicro Conference on Real-Time Systems*, July 2012, pp. 91–101.

[24] D. Wilkins, "The bathtub curve and product failure behavior, part one: The bathtub curve, infant mortality and burn-in," *Reliability Hotwire*, no. 21, 2002. [Online]. Available: http://www.weibull.com/hotwire/issue21/hottopics21.htm

[25] A. A. Balkema and L. de Haan, "Residual life time at great age," *The Annals of Probability*, vol. 2, no. 5, pp. 792–804, 1974. [Online]. Available: http://www.jstor.org/stable/2959306

[26] J. I. Pickands, "Statistical inference using extreme order statistics," *The Annals of Statistics*, vol. 3, no. 1, pp. 119–131, 1975. [Online]. Available: http://www.jstor.org/stable/2958083

[27] M. L. Hazelton, "Assessing log-concavity of multivariate densities," *Statistics and Probability Letters*, vol. 81, no. 1, pp. 121 – 125, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167715210002774

[28] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[29] S. Vilardell and Àlvar Pineda, *distTails: A Collection of Full Defined Distribution Tails*, 2019, r package version 0.1.2. [Online]. Available: https://CRAN.R-project.org/package=distTails

[30] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933. [Online]. Available: http://www.jstor.org/stable/91247

[31] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938. [Online]. Available: http://www.jstor.org/stable/2957648

[32] http://www.gaisler.com/index.php/products/processors/leon3, *Leon3 Processor*, Cobham Gaisler.

[33] C. Hernández, J. Abella, F. J. Cazorla, A. Bardizbanyan, J. Andersson, F. Cros, and F. Wartel, "Design and Implementation of a Time Predictable Processor: Evaluation With a Space Case Study," in *29th Euromicro Conference on Real-Time Systems (ECRTS 2017)*, ser. Leibniz International Proceedings in Informatics (LIPIcs), M. Bertogna, Ed., vol. 76. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017, pp. 16:1–16:23. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2017/7173

[34] J. del Castillo, D. M. Soler, and I. Serra, *ercv: Fitting Tails by the Empirical Residual Coefficient of Variation*, 2017, r package version 1.0.0. [Online]. Available: https://CRAN.R-project.org/package=ercv

[35] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial," *Biometrika*, vol. 26, no. 4, pp. 404–413, 1934. [Online]. Available: http://www.jstor.org/stable/2331986

[36] R. Fisher and L. Tippett, "Limiting forms of the frequency distribution of the largest or smallest member of a sample," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 24, no. 2, 1928.

[37] J. Hansen, S. Hissam, and G. A. Moreno, "Statistical-Based WCET Estimation and Validation," in *9th International Workshop on Worst-Case Execution Time Analysis (WCET'09)*, ser. OpenAccess Series in Informatics (OASIcs), N. Holsti, Ed., vol. 10. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2009, pp. 1–11, also published in print by Austrian Computer Society (OCG) with ISBN 978-3-85403-252-6. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2009/2291

[38] D. Griffin and A. Burns, "Realism in Statistical Analysis of Worst Case Execution Times," in *10th International Workshop on Worst-Case Execution Time Analysis (WCET 2010)*, ser. OpenAccess Series in Informatics (OASIcs), B. Lisper, Ed., vol. 15. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2010, pp. 44–53, the printed version of the WCET'10 proceedings are published by OCG (www.ocg.at) - ISBN 978-3-85403-268-7. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2010/2824

[39] L. Kosmidis, E. Quiñones, J. Abella, T. Vardanega, C. Hernandez, A. Gianarro, I. Broster, and F. J. Cazorla, "Fitting processor architectures for measurement-based probabilistic timing analysis," *Microprocessors*

and Microsystems, vol. 47, pp. 287 – 302, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0141933116300977

[40] Y. Lu, T. Nolte, I. Bate, and L. Cucu-Grosjean, "A new way about using statistical analysis of worst-case execution times," *SIGBED Rev.*, vol. 8, no. 3, pp. 11–14, Sep. 2011. [Online]. Available: http://doi.acm.org/10.1145/2038617.2038619

[41] A. Melani, E. Noulard, and L. Santinelli, "Learning from probabilities: Dependences within real-time systems," in *2013 IEEE 18th Conference on Emerging Technologies Factory Automation (ETFA)*, Sept 2013, pp. 1–8.

[42] S. Edgar and A. Burns, "Statistical analysis of wcet for scheduling," in *Proceedings 22nd IEEE Real-Time Systems Symposium (RTSS 2001) (Cat. No.01PR1420)*, Dec 2001, pp. 215–224.

[43] S. J. Gil, I. Bate, G. Lima, L. Santinelli, A. Gogonel, and L. Cucu-Grosjean, "Open challenges for probabilistic measurement-based worst-case execution time," *IEEE Embedded Systems Letters*, vol. 9, no. 3, pp. 69–72, Sept 2017.

[44] F. Guet, L. Santinelli, and J. Morio, "On the representativity of execution time measurements: Studying dependence and multi-mode tasks," in *17th International Workshop on Worst-Case Execution Time Analysis, WCET 2017, June 27, 2017, Dubrovnik, Croatia*, 2017, pp. 3:1–3:13. [Online]. Available: https://doi.org/10.4230/OASIcs.WCET.2017.3

[45] F. J. Cazorla, L. Kosmidis, E. Mezzetti, C. Hernandez, J. Abella, and T. V. of Padova, "Oprobabilistic worst-case timing analysis: Taxonomy and comprehensive survey," *ACM Computing Surveys*, vol. X, no. X, p. YYYY, 2019.

[46] L. Kosmidis et al., "Containing timing-related certification cost in automotive systems deploying complex hardware," in *DAC. (Best Paper Award)*, 2014.

[47] M. Dowle and A. Srinivasan, *data.table: Extension of 'data.frame'*, 2019, r package version 1.12.2. [Online]. Available: https://CRAN.R-project.org/package=data.table

[48] H. Wickham, *tidyverse: Easily Install and Load the 'Tidyverse'*, 2017, r package version 1.2.1. [Online]. Available: https://CRAN.R-project.org/package=tidyverse

[49] W. Chang, *extrafont: Tools for using fonts*, 2014, r package version 0.17. [Online]. Available: https://CRAN.R-project.org/package=extrafont

[50] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[51] S. Garnier, *viridis: Default Color Maps from 'matplotlib'*, 2018, r package version 0.5.1. [Online]. Available: https://CRAN.R-project.org/package=viridis