Applicability of Extreme Value Theory to the Execution Time Prediction of Programs on SoCs

Irina Fedotova, Bernd Krause and Eduard Siemens
Department of Electrical, Mechanical and Industrial Engineering,
Anhalt University of Applied Sciences, Bernburger Str. 55, 06366, Köthen,
Germany {irina.fedotova, bernd.krause, eduard.siemens}@hs-anahlt.de

Keywords: Extreme Value Theory, Worst-Case Execution Time, Probabilistic Timing Analysis, Timing Verification.

Abstract: This paper describes in detail the estimation algorithm of upper bound prediction of the time acquisition task. We use the specific hardware from ARM Cortex-A series and empirical approach of time values retrieval from the timer counter. The robust Measurement-Based Probabilistic Timing Analysis (MBPTA) method based on the Extreme Value Theory (EVT) has been used for experimental verification of the algorithm. The MBPTA method allows deriving a reliable and safe worst-case execution time (WCET) estimation based on the limited number of measurements on the target platform. However, it requires an appropriate complete set of statistical tests for verifying EVT applicability. In ongoing work, we intend to outline challenges behind EVT assumptions and parameter tuning for timing analysis, and provide more coherent approach for safe probabilistic WCET estimations in order to increase the confidence that timing constraints will be met.

1 INTRODUCTION

The timing validation process for real-time systems requires guarantees that the probability of the system failing to meet its timing constraints is below an acceptable threshold. Here the metric, which is used to prove that a task will complete its function in time, called the Worst Case Execution Times (WCET). However, due to diverse features of modern CPU, usually cycle-true simulation becomes infeasible and consequently imposes correspondent limitations to the aimed timing analysis. One possibility to get around this problem is the development of statistical methods, which could allow predicting the probability distribution of the circuit delay.

In general, the goal of all timing analysis is providing a safe upper bound of execution time for a particular task (Wilhelm, 2007). Nowadays, a number of different approaches pursued that goal, primary such as deterministic (DTA) and probabilistic approaches (PTA). The difference is mainly that the deterministic method produces a unique WCET estimate, while probabilistic - multiple WCET estimates with their respective probabilities. Each approach has its static (SDTA, SPTA) and measurement-based (MBDTA, MBPTA) variants. Classical static timing analysis (STA) operates on deterministic processor architectures and provide safe WCET estimates as they are proven to be the worst ones (Abella, 2014). STA uses the exact modeling of the system or a simulator, which is in practice for modern complex real-time systems a quite challenging task. In contrary to that, the measurement-based method provides an estimation based on the derived maximal and minimal observed execution times or their distributions. Therefore, WCET estimates retrieved by static methods adds a possible extra margin, whereas WCET estimates retrieved by measurement-based methods is simply the maximum value observed or assumed during measurements: $WCET_{measured} <= WCET_{exact} <= WCET_{static}$. Moreover, with a probabilistic hardware architecture and measurement-based approaches, it is possible to guarantee an accurate predicted WCET (probabilistic WCET or pWCET), what is taken upon itself by MBPTA approach.

For characterizing the worst-case, the MBPTA (Measurement-Based Timing Analyses) approach aims at modeling extreme execution times values, relying on measurements and the application of the Extreme Value Theory (EVT). The EVT in its turn deals with the extreme deviations from the median of probability distributions. It estimates the tails of distributions, where the worst case should lie. However, hardware systemic effects in real-time systems make EVT applicability difficult with regard

to its required theoretical hypotheses. Initially, in this paper, we focus on a particular hardware presented on the Atmel SAMA5D4 board, which is based on ARM Cortex-A series processors. It employs a cache of random-replacement policy, where the failure probability of 10^{-9} is facilitated (Altmeyer, 2015).

In our resent work (Fedotova, 2016), we have already outlined the main consequences of EVT assumptions and their correct interpretation. In the ongoing paper, we expand this research, in order to overcome rest difficulties with EVT checking all hypotheses for generalizing its applicability. Therefore, based on the previous related work (Abella, 2014), (Radojković, 2016), (Guet, 2016) and as well as on our empirical experiments, we suggest a consequent and systematic step-by-step method, in order to remove the existing ambiguities in applying EVT for providing probabilistic WCET estimations. Furthermore, we consider only the certain task of time acquisition on SoCs, unlike other works, where mostly known benchmark tools have been applied.

The rest of the paper is organized as follows: In Section 2 the related work on solving the WCET calculation problem is described. Section 3 introduces the problem of the probabilistic modeling and focuses on the theoretical aspects of the EVT applicability. In this section the main steps of the algorithm proposed in this paper is described as well. Section 4 describes the experiments and their setup on the used ARM Cortex A5 platform. Section 5, 6, 7 and 8 provide subsequent details and the requirements for EVT applicability. Particularly, the proof of fitting the target distribution, estimation parameters and obtaining WCET estimators. Finally, Section 9 concludes this work.

2 RELATED WORK

The first complete overview of modern methods for timing analysis of computer task has been done by R. Wilheim et al. (Wilheim, 2007). In this work, the classes of existing methods have been firstly presented. Particularly, the investigation of the correctness and precision of SPTA for systems that use a cache with an evict-on-miss random replacement policy have been described.

F. Cazorla et al. in (Cazorla, 2013) establish principles and requirements to EVT with the MBPTA method to derive WCET estimates. Thereby they addresses WCET problem by introducing randomization into the timing behavior of the system hardware and software. The work (Abella, 2014) presents comprehensive comparison among timing analysis techniques SDTA, SPTA and MBPTA. These and others works by these authors have been performed within the PROARTIS and PROXIMA projects for artificial random systems (random replacement policies in cache memories).

The work by F. Guet et al. (Guet, 2016) proposes a DIAGnostic tool, which applies the MBPTA method without human intervention. Depending on the certain theoretical hypotheses of the EVT, the logical work flow of the framework derives its pWCET estimate of traces of execution times. Also considering the DIAGnostic tool, K. Berezovskyi et al. investigate both methods of EVT “Block Maxima” in (Berezovskyi, 2014) and “Peak over Threshold” in (Berezovskyi, 2016) for Graphical Processor Units (GPUs). These works outline the particular features of each method. The main results have showed that hardware time-randomization is not essential for the applicability of EVT and can be applied even to some non-time-randomized systems as GPUs.

A statistical approach based on EVT theory has also been used for optimal performance analysis. Radojković et al. in (Radojković, 2016). Authors have presented an approach for finding and predicting the performance of the thread assignment in multi-core processors, using statistical inference.

However, aforementioned approaches give little information about the sequences of checking statistical hypotheses and making safe decision on their basis. In this work, adopting probabilistic analysis techniques, we intend to develop a more coherent analysis of the timing behavior on embedded platforms (in particular, considering the certain task of time acquisition).

3 THE PROBABILISTIC MODELING OF EXECUTION TIME

The measurement-based methods produce estimates (for parameters of some distributions) by executing the given task on the given hardware or on a simulator and measuring the execution time of the task or of its parts. In particular, MBPTA approaches are interested in modeling extreme execution times and characterizing the worst-case. The probabilistic theory that focuses on extreme values and large deviations from the average values is the Extreme Value Theory (EVT) (Coles, 2001). This section evaluates EVT theory by applying it to all measurements of time acquisition and outlines main steps to obtain reliable pWCET estimates - the worst possible distribution of task execution times.

3.1 EVT Applicability

The safety of the probabilistic worst-case estimates relates originally to the EVT applicability. The EVT theory estimates the probability of occurrence of extremely large values, which are known to be rare events. More precisely EVT predicts the distribution function for the maximal (or minimal) values of a set of n observations, which are modeled with random variables. The main result of EVT is provided in the Fisher-Tippett-Gnedenko theorem (Embrechts, 1996). The theorem characterizes the max-stable distribution functions, where $\{X_1, X_2, \ldots, X_n\}$ is a sequence of n independent and identically-distributed (i.i.d.) random variables and $M_n = max\{X_1, X_2, \ldots, X_n\}$. According to the theorem, if F is a non degenerate distribution function and there exists a sequence of pairs of real numbers (a_n, b_n) such that $a_n > 0$ and $lim_{n\to\infty} P([M_n - b_n]/a_n \leq x) = F(x)$, then F is called an extreme value distribution and belongs to one of the following three classes: either Fréchet, Gumbel, or Weibull.

In fact, these three distributions are combined in a single family of continuous CDFs, known as the generalized extreme value (GEV) distribution. Then GEV is characterized by three parameters: $\mu \in R$ - location parameter, $\sigma > 0$ - scale parameter and $\xi \in R$ - shape parameter. Depending on the shape parameter ξ, GEV has 3 types of distributions depicting the following three CDFs:

Type I, Gumbel ($\xi = 0$), when the underlying distribution has a nonheavy upper tail:

$$F(x;\mu,\sigma,\xi) = e^{-e^{-(x-\mu)/\sigma}}; \quad (1)$$

Type II, Fréchet ($\xi = \alpha^{-1} > 0$), when the underlying distributions has a heavy upper tail:

$$F(x;\mu,\sigma,\xi) = \begin{cases} e^{-y^{-\alpha}} & y > 0 \\ 1 & y \leq 0 \end{cases}; \quad (2)$$

Type III, "reversed" Weibull* ($\xi = -\alpha^{-1} < 0$), when the underlying distributions has a bounded upper tail:

$$F(x;\mu,\sigma,\xi) = \begin{cases} e^{-(-y)^{-\alpha}} & y > 0 \\ 1 & y \leq 0 \end{cases}; \quad (3)$$

where x is the total amount and y stands for the excess over the threshold u, with $y = x - u$. Figure 1 gives examples of Gumbel, Fréchet and Weibull distributions:

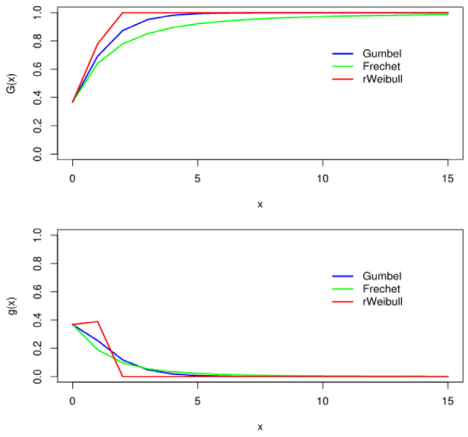


Figure 1: Examples of Gumbel, Fréchet and Weibull a) CDFs and b) PDFs with μ = 0 and σ = 1.

By estimating μ, σ, ξ parameters, we can prove the resulting GEV distribution: if the shape parameter $\xi = 0$, then the measured values in trace belong to a Gumbel distribution, which in most previous works (Cucu-Grosjean, 2012), (Hansen, 2009) has been assumed as applied to the pWCET distribution. Though there is no restriction on the values that ξ can take and resulting GEV distribution. Nevertheless, in order to be close to the accurate estimation of parameter, we intend to check all three distributions.

3.2 Selecting Extreme Values

Within the EVT context, there are two primary approaches to measure the extreme values: Block Maxima (BM) Models and The Peak over Threshold (POT). The approach of BM relies on deriving block maxima series. This is the traditional method, which comprises grouping the data into blocks, fitting the GEV distribution to the maxima of the blocks and estimating the risk measure from it. The second POT approach focuses on the observations, which exceed a given threshold. This is a more recent technique, which involves the following steps: select a threshold defining observations to include in modeling; calculate the exceedances; fit of the Generalized Pareto Distribution (GPD) to the exceedances and compute of the risk measure.

In fact, the Fisher-Tippet-Gnedenko theorem described above presents the EVT BM formulation where the tail distribution is the possible limit law characterizing the sequence of the maxima

* Within EVT the reverse (or negative) Weibull distribution is often referred to as the Weibull distribution, whereas the inverse Weibull is also known as type II or the Fréchet distribution

(Berezovskyi, 2014). Whereas in case of the BM approach the block size plays a central role, analogically the POT method models the law of the execution time peaks that exceed selected threshold. Nevertheless, the law of extreme execution times and the BM are closely linked to the law of peaks above the thresholds. Since the same value of ξ is shared, the equivalence of the distribution laws composing both the GEV and GPD distributions can be followed (Berezovskyi, 2016). The Pickands-Balkema-de Haan theorem presents the formulation of POT method (Balkema, 1974). Accordingly to it, for a large class of underlying distribution function F (which satisfies the conditions of Fisher-Tippet-Gnedenko Theorem) the conditional excess distribution function $F_u(x)$,

$$F_u(x) = F_u(y+u) = P(X-u \le y | X > u) \quad (4)$$
$$= \frac{F(x)-F(u)}{1-F(u)} \quad for\ 0 \le y \le x_0 - u,$$

is well approximated by the Generalized Pareto Distribution $G_{\xi,\sigma_u}(x)$:

$$\lim_{u \to x_0} \sup_{0 \le y < x_0 - u} |F_u(x) - G_{\xi,\sigma_u}(x)| = 0$$

$$G_{\xi,\sigma_u}(x) = \begin{cases} 1 - (1 + \xi \frac{y}{\sigma_u})^{-1/\xi} & \xi \neq 0 \\ 1 - e^{-y/\sigma_u} & \xi = 0 \end{cases} \quad (5)$$

where x_0 is either finite or infinite right endpoint of the underlying distribution F; $y \ge 0$ when $\xi \ge 0$, and $0 \le y \le -\sigma_u/\xi$ when $\xi < 0$; and $\sigma_u = \sigma_{u_0} + \xi(u - u_0)$;

As in recent studies have been followed, the POT is preferred over the BM, because data are used more efficiently, though the evident disadvantage is the selection of the suitable threshold value. Moreover, in the single-path case the POT appears to be more accurate (with respect to the measurements), but the increase of the threshold u can result into more pessimistic pWCET estimations. Further, we use the POT method to estimate the cost of time acquisition, based on the measured cost of the sample. However, the complexity due to threshold selection and its impact to the resulting pWCET has to be considered.

3.3 WCET Estimation Algorithm

The following algorithm for WCET estimation is suggested according to probabilistic theory described above:

Step 1. Selecting extreme values. The objective of this step is to collect, from the original distribution, the values, which fit into the tail, and hence can be modeled with the GEV distribution. Further, the POT method has been chosen to estimate extremes:

Step 1.1. Threshold choice. With the help of graphical diagnostic: mean residual life plot and parameter stability plot, choice the best fitted threshold.

Step 1.2. Retrieve the new sample data: filter values which are above the threshold.

Step 2. Fitting the GEV distribution: Gumbel, Fréchet and Weibull types. To ensure that the sample data correctly matches the distribution we fit, the certain goodness-of-fit tests as well as Chi-square, QQ-plot have been used. If none of three types distributions fits, then going back to Step 1.1 and increase the threshold value.

Step 3. Estimate the remaining parameters of fitted distribution: μ, σ and ξ.

Step 4. Verification of EVT hypothesis of independence and identical distribution. If both are verified, then the EVT distribution tail projection can be considered as a safe and good pWCET estimate.

Step 4.1. Checking that the data are identically distributed.

Step 4.2. Prove that samples are independent. That is ensured by a combination of hardware with suitable randomization properties.

Step 5. Return WCET estimation based on μ, σ and ξ parameter.

4 EXPERIMENTAL SETUP

Within MBPTA approach, complete runs of the test are made on the target hardware. For these experiments, we have collected timestamps of CPU cycle counter on the Atmel SamaA5D4 board. This board uses one ARM Cortex-A5 600 MHz core, which belongs to the ARMv7-A architecture generation. The common problem of most processors in SoCs is, that CPU cycle counter is not directly available from user-space. Thereby, the investigations of timing capabilities are being performed within the high performance *HighPerTimer* library (Fedotova, 2013). The main idea behind the *HighPerTimer* library is to simplify the timestamps acquisition process from the main cycle counter of different processors. During the library initialization step, a specific time register is assigned to the main library time source. The time counter has the channel size of 32-bit width and frequency 11 MHz. Thus, it wraps around in every 6.5 minutes. The said library provides means for correct dealing with such wrap-arounds and providing a global 64-bit fast ticks counter independently from the underlying timing hardware. A special device

driver within the library, which is loaded as a kernel module beforehand, enables proper operations with timers from the user-space. The main timing mechanism, which in its turn supports the procedure of handling overflows is processed by the user space library avoiding any system calls.

The platform is running with the standard Linux kernel of version 4.4.11 and the measured process is scheduled by the Normal scheduling policy, which is set by default. Within the context of this work, we estimate the timer cost, setting two consecutive timers of *HighPerTimer* library and calculating the time difference between them (further x_i.). The representation of complementary cumulative distribution function (CCDF) of the trace is shown in Figure 2 and basic statistics in Table 1.

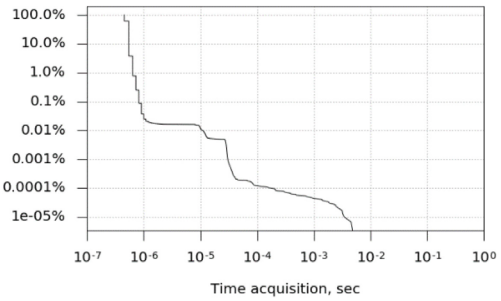


Figure 2: CCDF representation of time execution.

Table 1: Statistical properties of the original data set.

Number of samples	$30*10^6$
Trace length	27 min
Mean execution time	5.71 cycles = 0.519 µsec
Std. deviation	24.384 cycles = 2.217 µsec
Max value	65943 cycles = 5994.7 µsec
Min value	5 cycles = 0.455 µsec

The estimation part of a representative trace has been taken 27 minute and as can be seen from the graph, the distribution peaks near the mean and falls with rapidly decreasing probability density below 1 µsec.

5 GRAPHICAL DIAGNOSTICS FOR THE OPTIMAL THRESHOLD

The choice of an appropriate threshold u requires a compromise between precision and bias. If the threshold is too low, then the results will tend to be more certain. On the other hand, the analysis will only become practically valid, when the threshold is sufficiently high. Therefore, the goal is to find such a lowest possible threshold, that the extreme value model provides a reasonable fit to exceedances of it. Two graphical tools can be used for identifying an appropriate threshold for modeling extremes via the GPD: Mean residual life plot and Parameter stability plot (Scarrott, 2012).

a) Mean residual life plot.

In the mean residual life plot, for a range of candidate values for u the corresponding mean threshold excess has to be identified. Then this mean threshold excess is plotted against u. The plot should be linear above the threshold u_0 at which the GPD model becomes valid. On the Figure 3, the blue lines correspond to the lower and upper confidence limits respectively. The purpose is to find the lowest threshold where the plot is nearly linear, taking into account the 95% confidence interval. Though interpretation of these plots can be subjective, linearity in Figure 3 might be suggested above $u_0 \approx 2e\text{-}04\ sec$, beyond which it is approximately linear until $u \approx 2.5e\text{-}04\ sec$, whereupon it decreases sharply. These limits are dashed red lines on the plot. This way, the minimum and maximum possible thresholds, at which the model can be fitted have been firstly suggested as $u_{min} \approx 2e\text{-}04$ and $u_{max} \approx 2.5e\text{-}04$.

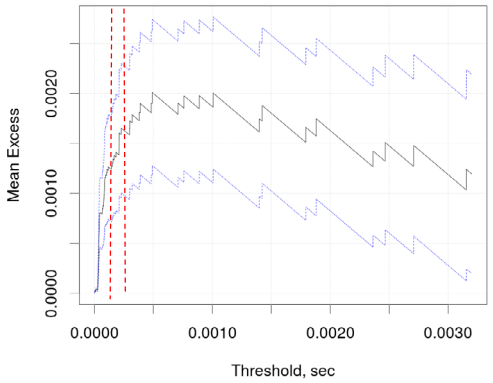


Figure 3: The empirical mean residual life plot.

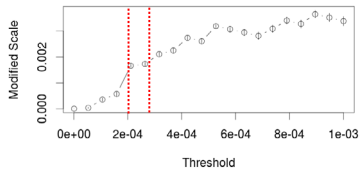
b) Parameter stability plot.

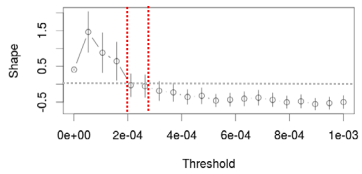
For the next step, two parameter stability plots showing maximum likelihood estimates, confidence intervals of the shape and modified scale parameters over a range of thresholds are produced. Figure 4 represents plots from fitting the GPD and point process models to these data. Denoting the value of the generalized Pareto scale parameter σ_u for a threshold from $u > u_0$ in (5), the scale parameter changes with u unless $\xi = 0$. Thereby, we can better

express it as a constant scale parameter with respect to u (Coles, 2001):

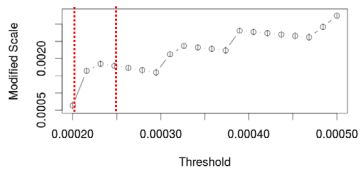
$$\sigma^* = \sigma_u + \xi u, \quad (6)$$

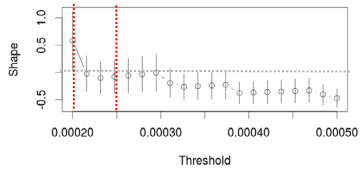
Consequently, estimated of σ^* and ξ should be constant above u_0 or at least stable after sampling errors. Therefore firstly, comparing the parameter stability plot for the whole range of samples in Figure 4(a) with the mean residual life plot in Figure 3, the retrieved possible interval of u_{min} and u_{max} can be confirmed. At least for the case of Gumbel distribution with its shape parameter $\xi = 0$, the desired threshold is likely within this range. Figure 4(b) shows more accurately the desired range, where the dependence of ξ parameter can be better observed.





a)





b)

Figure 4: Parameter stability plot a) for the whole trace and b) for the range with u_{min} = 2e-04 sec.

Though the threshold stability plots also does not provide very firm conclusions, together with the mean residual life plot, inconsistencies can be good observed. It lies between the estimated shape parameter at this level and higher thresholds around *u=2.3e-04* sec. Therefore, at the current stage it can be concluded that this is the best choice of threshold, which allows retrieving 24 extremes from the trace.

6 FITTING THE GPD DISTRIBUTION AND ESTIMATION THE PARAMETERS

Having determined the threshold value, the parameters of the GPD can be further estimated by maximum likelihood (MLE). In the following section we have tested the whole family of GPD distributions. The appropriate Goodness-of-fit statistics for Gumbel, Fréchet and Weibull distributions have been obtained and presented in Table 2. Each test is essentially a goodness of fit test and compares observed data to quantiles of the specified distribution. The null hypothesis for each test versus alternative is:

H_0: data follow an assigned distribution;

H_1: data do not follow an assigned distribution.

The resulting value is then checked against the following statistics to see if it is significant:

– the critical value from Chi-square test. Since we deal with the discrete data, the Chi-square test has been chosen against Kolmogorov Smirnov test. From the Chi-Square table we can find the critical Chi Square value for a level of significance p, which represents the probability that a Chi Square distributed random variable will exceed that critical value. Typically a match at the $p = 0.05$ is considered acceptable.

– the Bayesian (BIC) and Akaike (AIC) information criterion of (Burnham, 2004). AIC tries to select the model that most adequately describes an unknown one. Conversely, BIC aims to find the true model among the set of candidates. When comparing models fitted by maximum likelihood to the same data, the smaller the AIC or BIC, the better the fit.

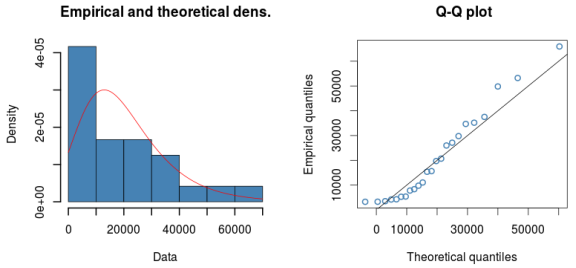
– a QQ-plot (quantile plot). This is a plot of the empirical quantile values of observed data against the quantiles of the standard form of a target distribution. The slope and the intercept of the best-fit line through these points can be used as estimators ξ and σ parameters, respectively. A straight diagonal line of data points from the bottom left to the top right of the

Table 2: GPD parameters.

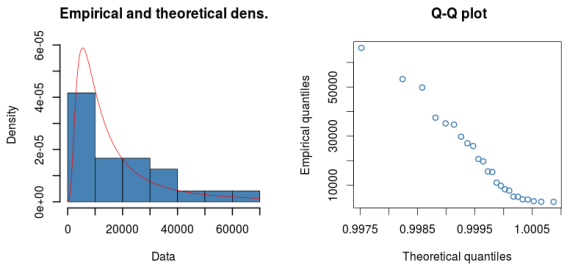
	p-value	AIC	BIC	σ	μ	ξ
Gumbel	0.346	534.240	536.596	12247.56	12968.35	0
Fréchet	0.065	531.211	534.745	10542.52	-1539.657	1.351

plot indicates that an exponential distribution is a relatively good fit to the tail.

Firstly, Chi-squared p-value for the Gumbel and Fréchet data ($p\text{-}value_G = 0.346$ and $p\text{-}value_F = 0.065$ respectively) > 0.05, hence both hypotheses can't be rejected. Secondly, the goodness-of-fit criteria for the Fréchet distribution is a bit less than for the Gumbel distribution: $AIC_F = 531.211 < AIC_G = 534.240$ and $BIC_F = 534.745 < BIC_G = 536.596$. Therefore, from the statistical tests the hypotheses that data fits Fréchet distribution is slightly preferred against the one, which checks Gumbel distribution. However, since the difference of statistics results is not significantly differed, it makes sense to retrieve the WCET estimate for both cases.



a)



b)

Figure 5: The histogram against fitted density functions and theoretical quantiles against empirical ones of a) Gumbel and b) Fréchet distribution.

7 VERIFICATION OF EVT HYPOTHESIS

In order to apply the EVT, three hypotheses are required to verify: i) independent and ii) identically distributed execution time measurements from iii) a distribution, which belongs to the Maximum Domain of Attraction of the GEV with a shape parameter ξ (Guet, 2016). These proofs provide reliable and safe pWCET estimates. In the giving chapter, we intend to check the first two hypothesis. As follows from the definition, the sequence of random variables is independent and identically distributed if each random variable has the same probability distribution as the others and all are mutually independent (Coles, 2001), (Burnham, 2004), (Feller, 1996).

a) Identical Distribution.

It is worth to note that in the given application of EVT, the rule of identically distributed is obeyed since the analysis models the behavior of the system in the same execution context using the same set of parameters, including initial hardware and software state (Cazorla, 2013). However, to provide enough level of reliability, the property of identical distributed values are verified by the two-sample Kolmogorov-Smirnov (KS) test. The Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution functions of two samples. For the given experiment the null (H_0) and alternative (H_1) hypotheses are:

H_0: the both samples are identically distributed;

H_1: the samples are not identically distributed.

The test is performed by dividing the trace into subsets in order to verify if they have the same distribution function. By randomly taking elements from the original sample, three subsets of 100, 500 and 1000 values have been created. This ensures that the smaller samples maintain the same statistical properties as the original (Cucu-Grosjean, 2012), (Burnham, 2004), (Feller, 1996). Table 3 represents the p-value obtained by applying the KS test to the execution times. *P-value* is the probability of finding a situation more extreme than what in the data, assuming that $a_n = b_n$. The smaller this number is, the less likely that $a_n = b_n$ is true. *D value* of the KS test statistic means the maximum difference between the a_n & b_n probability mass function. The rule to accept H_0 is $p\text{-}value > 0.05$ - the predetermined significance level. Accordingly, from Table 3, the null hypothesis cannot be rejected, which allow concluding that samples are identically distributed.

Table 3: Statistics for identical distribution test.

m	D value	p-value
100	0.1	0.6994
500	0.012	1
1000	0.053	0.125

b) Independence.

From the definition, two random variables are considered to be independent if they describe two events such that the occurrence of one event does not have any impact on the occurrence of the other event (Coles, 2001), (Burnham, 2004), (Feller, 1996). To prove those properties, the Runs test (or Wald-Wolfowitz test) (Feller, 1996), (Wald, 1940), is used. In this context, a term "run" is a sequence of identical responses. The null and alternative hypotheses are:

H_0: elements of the sequence are mutually independent;

H_1: elements of the sequence are not mutually independent.

The following steps have to be accomplished to apply the Runs test:

Step 1: compute the sequential differences $d_i - d_{i-1}$, where positive values is related to increasing values and negative to a decreasing ones.

Step 2: compute the expectation of the number of runs $E(R) = 2mn/N$, where N is the total sample size, m is the number of positive values, and n is the number of negative ones.

Step 3: compute the variance of the number of runs $V(R) = 2mn(2mn - N)/(N^2(N-1))$. The minimum value of R is always 2. The maximum value is given by $2Min(m, n)-t$, where $t = 1$ for $m = n$, and $t = 0$ if not.

Step 4: estimate the test statistic $Z=(r-E(R))/\sqrt{V(R)}$. In Table 4, Z value and *p-value* are compared for a significance level $\alpha = 5\%$. At the given level, *Z-value* with an absolute value greater than 1.96 indicates non-randomness so the null hypothesis is rejected. Additionally, the rule to accept H_0 is if *p-value* is more than 0.05.

Since the $p\text{-value} = 0.247 > 0.05$ and $Z\ value = 0.684 < 1.96$, the hypothesis that each element in the sequence is independently drawn from the same distribution is accepted.

Table 4: Statistics for independence test.

V(R)	Z value	p-value
19.8	0.684	0.247

8 ESTIMATION PROBABILISTIC PWCET

The final step is to use the computed and verified GPD parameters and the exceedance probability of failure p to estimate the WCET. In fact, WCET thresholds are defined depending on the failue probability p such that $p = P(WCET_{safe} > WCET_{exact})$. In standard statistical language, this is a quantile estimate or for instance in finance, it is often referred as Value-at-risk (VaR) for measuring of market risk. Considering these application for our case, the WCET estimation is then derived on the basis of estimated parameters $\hat{\xi}$ and $\hat{\sigma}$ as following (Embrechts, 1996):

$$WCET = \begin{cases} u + \frac{\hat{\sigma}}{\hat{\xi}}((\frac{n}{k}p)^{-\hat{\xi}} - 1) \; if \; \hat{\xi} > 0 \\ u - \hat{\sigma} log(\frac{n}{k}p) \qquad if \; \hat{\xi} = 0 \end{cases} \quad (7)$$

where k - the number of peaks over the threshold standing for measurements that belong to the tail distribution. The initial probability of failure p defined as the likelihood the execution of a job exceeds its WCET for the current mode when previous jobs have not exceeded it. Different works on probabilistic WCET have claimed that values of p could typically be 10^{-16}, 10^{-9} or 10^{-4}. For our investigations, the failure probability at the level $10^{-9} < p < 10^{-7}$ (from hazardous class) provided by PED certification (Hsing, 1991), (ARP4761, 2001) has been chosen. PED certification is applied in the flight control system using portable electronic devices (PEDs). Further, Table 5 gives the modeling results of the extreme execution times and Figure 7, showing the distribution convergence for Gumbel and Fréchet scenarios.

In Figure 6 the x-axis shows the pWCET estimation and the y-axis shows the associated probabilities. The challenge of risk assessment is to assess the value that for each of the activities is not

Table 5: EVT Results for the Hazardous Class of p Considering GPD.

	σ	**μ**	**ξ**	**(WCET ; 10^{-7})**	**(WCET ; 10^{-8})**	**(WCET ; 10^{-9})**
Gumbel	12247.56	12968.35	0	27968.09 cycles = 2.54 msec	56169.13 cycles = 5.11 msec	84370.18 cycles = 7.67 msec
Fréchet	10542.529	-1539.657	1.351	124185.2 cycles = 11.3 msec	2898825 cycles = 263 msec	65126941 cycles = 5.921 sec

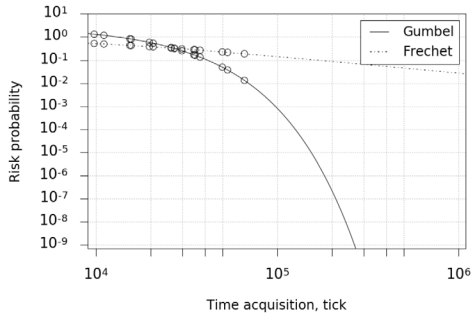


Figure 6: Estimate of upper bounds for the Gumbel and Frechet distribution.

able to be measured accurately. So we use a CCDF of General Pareto Distribution to predict these potential values. Finally, all tests are diagnosed as reliable regarding the EVT applicability and the pWCET estimate has been derived.

However, according to the Table 5 the WCET estimates for Gumbel and Fréchet distribution differ significantly. In fact, too pessimistic results of Fréchet are quite tight and practically less useful. On the other hand, the Gumbel distribution converges to 0 faster than the Fréchet one and it decrease the pessimism of the WCET thresholds. Nevertheless, having the access to the true values and capacity to run the experiment including the target risk probabilities, we can estimate the real max value and predicted WCET estimates. According to the Table 1, the max value on the examined range is 0.005 sec and whereas predicated upper bound using parameters for Gumbel distribution for failure probability $p = 10^{-9}$ gives 0.007 sec, which fits the assumption. For the case of Fréchet parameters, the estimate of 5.921 sec is obviously too pessimistic and less useful.

9 CONCLUSION

The contributions of this paper are: (i) proving optimality of the EVT theorems and verifying their applicability on the certain embedded platform ARM Cortex-A series of processors; (ii) introducing an approach of graphical diagnostic for selecting extreme values; (iii) considering several cases of GPD distributions for choosing reliable WCET estimation. Our results show that, for failure probability levels of 10^{-9} the single-path technique for time acquisition provides less pessimistic pWCET estimations about 7.67 msec using parameters for Gumbel distribution. This estimation can be taken as acceptable and considered during designing of time-critical applications on such SoCs. Therefore, the ongoing paper provides a useful guide of how to predict the upper bounds for the embedded single-core processor architecture. The future work should continue investigation of cache memory effects or impacts of scheduling tasks by Linux kernel. This can afford the improvement of the dependence metrics, reduce the pessimism of the pWCET estimation and as a result, make ARM processor more time-predictable.

REFERENCES

Wilhelm, R., Engblom, J., Ermedahl, A., Holsti, N., Thesing, S., Whalley, D., Bernat, G., Ferdinand, C., Heckmann R., Mitra, T., Mueller, F., Puaut, I., Puschner, P., Staschulat, J., Stenström, P., 2007, 'The Worst-Case Execution Time Problem — Overview of Methods and Survey of Tools', *ACM Trans. Embed. Comput. Syst.*, pp. 36–53.

Abella, J., Hardy, D., Puaut, I., Quinones, E., Cazorla, FJ., 2014, 'On the Comparison of Deterministic and Probabilistic WCET Estimation Techniques', *Proc. of the 26th Euromicro Conference on Real-Time Systems (ECRTS'14)*, Madrid, pp. 266–275.

Fedotova, I., Krause, B., Siemens E., 2017, 'Upper Bounds Prediction of the Execution Time of Programs Running on ARM Cortex-A Systems', Submitted at *IFIP AICT (Advances in Information and Communication Technology)* in press.

Altmeyer, S., Cucu-Grosjean, L., Davis, RI., 2015, 'Static probabilistic timing analysis for real-time systems using random replacement caches', *Real-Time Syst.*, vol. 51, no. 1, pp. 77–123.

Radojković, P., Carpenter, PM., Moretó, M., Čakarević, V., Verdú, J., Pajuelo, A., Cazorla, FJ., Nemirovsky, M., Valero, M., 2016, 'Thread Assignment in Multicore/Multithreaded Processors: A Statistical Approach', *IEEE Trans. Comput.*, vol. 65, no. 1, pp. 256–269.

Guet, F., Morio, J., Santinelli, L., 2016, 'On the Reliability of the Probabilistic Worst-Case Execution Time Estimates', *the 8th European Congress on Embedded Real Time Software and Systems (ERTS 2016)*, Toulouse, pp. 758–767.

Cazorla, F., Vardanega, T., Quinones, E., Abella, J., 2013, 'Upper-bounding Program Execution Time with Extreme Value Theory', *Proc. WCET workshop*, Paris, pp. 64-76.

Berezovskyi, K., Santinelli, L., Bletsas, K., Tovar, E., 2014, 'WCET Measurement-based and Extreme Value Theory Characterisation of CUDA Kernels', *Proc. of the 22nd International Conference on Real-Time Networks and Systems*, New York, , pp. 279-288.

Berezovskyi, K., Guet, F., Santinelli, L., Bletsas, K., Tovar, E., 2016, 'Measurement-Based Probabilistic Timing Analysis for Graphics Processor Units', Proc. Of *29th International Conference Architecture of Computing Systems – ARCS 2016*, Nuremberg, pp. 223-236.

Coles, S., 2001 ed, *An Introduction to Statistical Modeling of Extreme Values*. Springer.

Cucu-Grosjean, L., Santinelli, L., Houston, M., Lo, C., Vardanega, T., Kosmidis, L., Abella, J., Mezzetti, E., Quinones, E., Cazorla, JF., 2012, 'Measurement-Based Probabilistic Timing Analysis for Multi-path Programs', *Proc. of the 24nd Euromicro Conference on Real-Time Systems (ECRTS)*, pp. 91-101.

Hansen, J., Hissam, SA., Moreno, GA., 2009, 'Statistical-Based WCET Estimation and Validation', *Proc. of the 9th Intl. Workshop on Worst-Case Execution Time Analysis (WCET)*, pp. 123-133.

Embrechts, P., Klueuppelberg, C., Mikosch, T., 1996, *Modelling Extremal Events for Insurance and Finance*, Springer, Berlin.

Balkema A., Haan L., 1974, 'Residual Life Time at Great Age', *Annals of Probability*, vol. 2, no. 5, pp-749-791.

Fedotova I., Siemens E., Hu H., 2013, 'A High-precision Time Handling Library', *Journal of Communication and Computer*, vol. 10, pp. 1076–1086.

Scarrott, C., MacDonald A., 2012, 'A Review of Extreme Value Threshold Estimation and Uncertainty Quantification', *REVSTAT - Statistical Journal*, vol. 10, no. 1, pp. 33–60.

Burnham, KP., Anderson, DR., 2004, 'Multimodel Inference: Understanding AIC and BIC in Model Selection', *Sociological Methods & Research*, vol. 33, no. 2, pp. 261–304.

Wald, A., Wolfowitz, J., 1940, 'On a Test Whether Two Samples are from the Same Population', *Annals of Mathematical Statistics*, vol. 11, no. 2, pp. 147–162.

Hsing, T., 1991, 'On Tail Index Estimation Using Dependent Data', *Annals of Statistics*, vol. 19, no. 3, pp. 1547–1569.

ARP4761, 2001, '*Guidelines and methods for conducting the safety assessment process on civil airborne systems and equipment*'.