EVT-based Worst Case Delay Estimation Under Process Variation

Charalampos Antoniadis, Dimitrios Garyfallou, Nestor Evmorfopoulos, and Georgios Stamoulis
Dept. of Electrical & Computer Engineering, University of Thessaly, Volos, Greece
{haadonia, digaryfa, nestevmo, georges}@e-ce.uth.gr

***Abstract*—Manufacturing process variation in sub-20nm processes has introduced ever increasing overhead in Static Timing Analysis (STA) in order to guarantee the reliable operation of the circuit. Chip designers apply corner-based analysis and add guard-bands to design parameters in order to take into account the impact of process variation on timing. However, the aforementioned techniques are either too slow as the number of design parameters proliferates with the integration of more components into a chip or inaccurate due to the assumption that the worst case delay resides at the corners of design parameters. In this paper, we present a novel statistical methodology, which relies on Extreme Value Theory (EVT), to estimate the worst case delay of VLSI circuits under variations in gate/interconnect parameters. Despite the previous statistical approaches toward maximum delay estimation, our methodology can be applied regardless of the underlying gate/interconnect delay model or any assumption about the distribution of the Arrival Time (AT) at every circuit node, making it very appealing for integration to any level of timing analysis abstraction (from spice-to-gate level) and provide fast yet accurate results. Experimental results on ISCAS85/ISCAS89 circuits show that the estimated maximum AT at the Primary Outputs (POs) can be within 5% of the true maximum AT, at the cost of a few thousand Monte Carlo simulations.**

I. Introduction

The advent of the aggressive technology scaling era has introduced extensive spreads in the transistor and interconnect parameters. Gate and interconnect delay, which is a function of the aforementioned parameters, has to be considered random, distributed between a minimum and a maximum value. As a result, chip designers have to take into account process variation so that they can ensure reliable operation of the circuit. We can distinguish two approaches to verify timing under variation. The first one and most often used-in industry is corner-based analysis. In the corner-based methodology the circuit is simulated at the corners of the design parameters where designers expect to find the worst-case delay. In the second method, timing verification is performed through *Statistical Static Timing Analysis* (SSTA), looking for the distributions of ATs at the *Primary Outputs* (POs). These distributions then, provide useful information that enable the estimation of worst-case AT. SSTA can be performed either by distribution propagation from any *Primary Input* (PI) to any PO or by *Monte Carlo* (MC) simulations. The former method is faster, but leads to less accurate results, than the latter one.

Although previous works toward worst case delay estimation are attractive and provide meaningful insights into the problem, they are either too slow for large designs or inaccurate, as they are based on simplistic assumptions about the underlying delay models for the interconnect/gates or the propagated distributions across the circuit nodes when SSTA is performed. More specifically, [1] approximates the result of the non-linear MAX operation between two random variables normally distributed with a normal distribution. However, they do not test the effectiveness of their proposed approximation on a convincing set of benchmarks. In any case the assumption they make is not supported by the stochastic process theory and we expect their approximation to deviate even more from the true distribution when complex gates with more than two inputs are considered. Furthermore, [2] and [3] model the delay response with a *Response Surface Modeling* (RSM) method and obtain the worst delay based on that model. However, due to an unprecedented increase in transistor and interconnect complexity (large number of metal layers) [4] RSM would require prohibitive number of simulations to cover all the dimensions of the problem in the design of the experiments. The authors in [5] focus only on one stage and derive the worst case delay condition studying different interconnect structures and gate drive strengths. They provide useful guidelines for the selection of the capacitance and resistance values for an interconnect that results in the worst-case delay for the stage. However, they do not comment on how the worst case delay condition can be extended to the whole circuit, as it is infeasible to extract a global worst case condition due to the existing inter-dependencies between the delay of a gate/interconnect and the parameters of the following gates/interconnect on a path.

To this end, in this paper we present a novel statistical methodology, which relies on EVT to estimate the worst AT at the POs of VLSI circuits, under variations on parameters that determine gate and interconnect delay. The contribution of our work is that the proposed methodology can provide fast yet accurate results irrespective of the timing models or any assumption about the distributions of AT at the circuit nodes.

The rest of the paper is organized as follows. Section II provides a brief introduction to EVT and presents how an upper end point of a bounded random variable can be estimated, exploiting elements from EVT. Section III explains why it is infeasible to extract a global worst case condition for the parameters that affect the delay under process variation apriori. Section IV presents our methodology for the worst case delay estimation. Section V comments on the results of the proposed methodology applied on the ISCAS benchmarks. Finally, conclusions are drawn in Section VI.

II. Theoretical Background

A. Modeling Extreme and Rare events

Extreme value theory is a branch of probability theory that focuses on the study of extreme and rare events. There are two possible methods of modeling extreme statistics on the basis of a random sample $\mathbf{X}_n = \{X_1, X_2, \ldots, X_n\}$[1], where X_is are *independent, identically distributed* (i.i.d) random variables from the *cumulative density function* (cdf) F, namely the *Block Maxima* method and the *Peak Over Threshold* method. The first one divides the data sample $\mathbf{X}_n$ into l blocks of size m and then holds the maximum from each block, making a new sample of *maximas* $\mathbf{M}_l = \{M_1, M_2, \ldots, M_l\}$. The sample of maxima follows a distribution with cdf that is given by [6]:

$$F_M = P(X_1 \leq x, \ldots, X_m \leq x) = \prod_{i=1}^{m} P(X_i \leq x) = F^m(x) \tag{1}$$

[1]Bold letters denote vector variables, while non-bold letters denote scalar variables.

978-3-9819263-0-9/DATE18/©2018 EDAA

The other method takes the k largest values from the data sample $\mathbf{X}_n$ that exceed a predetermined high threshold u and forms a separate sample of *exceedances* ($\mathbf{X}_{ex}$). Again, assume that the sample of exceedances consists of i.i.d random variables from the cdf F. The sample of exceedances follows a distribution with (conditional) cdf that is given by [6]:

$$F_Z(z) = P(X - u \leq z | X > u) = \frac{F(z+u) - F(u)}{1 - F(u)} \quad (2)$$

B. EVT: Limiting Distributions

We first need to define the concept of the *upper end point*, which plays a central role in the prediction of the worst case delay.

Definition 1: The upper (or right) end point $\omega(F)$ of cdf $F(x)$ is defined as the upper bound of the support of $F(x)$:

$$\omega(F) = sup\{x : F(x) < 1\} \quad (3)$$

The upper end point represents the maximum value that the associated random variable can acquire and becomes $\omega(F) = F^{-1}(1)$ if the random variable is bounded or $\omega(F) = +\infty$ in the opposite case.

The two fundamental theorems, upon which EVT relies, designate the limiting distributions of maxima sample defined in eq. (1) when $m \to \infty$ and the limiting distribution of exceedances over a threshold defined in eq. (2) when $u \to \omega(F)$.

Theorem 1 *(Fisher-Tippet [7])*: Sample Maxima cdf (F_M), for given normalizing constants, a_m, b_m converges to the *Generalized Extreme Value* (GEV) as m tends to infinity:

$$\lim_{m \to \infty} F_M(a_m x + b_m) \to H_\xi = e^{-(1-\xi x)^{\xi^{-1}}} \quad (4)$$

where ξ is a parameter that determines the shape of H and depends on $F(x)$.

H can be classified, with respect to the shape parameter ξ, into one of the following cdfs:

Frechet:

$$H_{\xi<0}(x) = \begin{cases} 0, & x \leq \mu_m \\ e^{-(\frac{x-\mu_m}{\sigma_m})^{-\xi^{-1}}}, & x > \mu_m \end{cases}$$

where $\mu_m = 0$ and $\sigma_m = F^{-1}\Big(1 - \frac{1}{m}\Big)$.

Weibull:

$$H_{\xi>0}(x) = \begin{cases} e^{-(-(\frac{x-\mu_m}{\sigma_m})^{\xi^{-1}})}, & x \leq \mu_m \\ 1, & x > \mu_m \end{cases} \quad (5)$$

where $\mu_m = \omega(F)$ and $\sigma_m = \omega(F) - F^{-1}\Big(1 - \frac{1}{m}\Big)$.

Gumbel:

$$H_{\xi \to 0}(x) = e^{-e^{-\frac{x-\mu_m}{\sigma_m}}}, \quad x \in \Re \quad (6)$$

where $\mu_m = F^{-1}\Big(1 - \frac{1}{m}\Big)$ and $\sigma_m = m \int_{F^{-1}(1-\frac{1}{m})}^{\omega(F)} (1 - F(y))dy$.

Theorem 2 *(Balkema and de Haan [8] and Pickands [9])*: Exceedances over threshold (X_{ex}) (conditional) cdf, for a given scale factor b_u, converges to the *Generalized Pareto* (GP) cdf as u tends to $\omega(F)$:

$$\lim_{u \to \omega(F)} F_Z(\frac{z}{b_u}) \to GP_\xi(z) = 1 - (1 - \xi z)^{\xi^{-1}} \quad (7)$$

where ξ is a parameter that determines the shape of GP and depends on $F(x)$.

Depending on ξ, GP_ξ belongs to one of the following distribution families:

Pareto:

$$GP_{\xi<0}(x) = \begin{cases} 0, & x \leq u \\ 1 - (\frac{(x-u)+\sigma_u}{\sigma_u})^{\xi^{-1}}, & x > u \end{cases}$$

where $\sigma_u = u$.

Beta:

$$GP_{\xi>0}(x) = \begin{cases} 1 - (-\frac{(x-u)-\sigma_u}{\sigma_u})^{\xi^{-1}}, & u < x \leq u + \sigma_u \\ 1, & x > u + \sigma_u \end{cases} \quad (8)$$

where $\sigma_u = \omega(F) - u$.

Exponential:

$$GP_{\xi \to 0}(x) = \begin{cases} 0, & x < u \\ 1 - e^{-\frac{x-u}{\sigma_u}}, & x \geq u \end{cases} \quad (9)$$

where $\lim_{u \to \omega(F)} \frac{\sigma_u}{\xi(\omega(F)-u)} = 1$ when $\xi > 0$ and $\lim_{u \to \omega(F)} \frac{\sigma_u}{-\xi u} = 1$ when $\xi < 0$.

C. Estimation of a finite upper end point $\omega(F)$

An important fact derived from the limiting cdfs in eq.(5, 8) is that a parent cdf with an infinite upper end point ($\omega(F) = +\infty$) can only have an extreme value distribution with $\xi < 0$, whereas a cdf with a finite upper end point ($\omega(F) < +\infty$) suggests an extreme value distribution with $\xi > 0$. Notice that Weibull and Beta cdfs reach 1 for all x values greater than a finite threshold. However, if $\xi \to 0^+$ ($\xi \downarrow 0$), then $\omega(F)$ cannot be efficiently or accurately estimated through the previous set of distributions and parameters constructed for the general case $\xi > 0$ [10] and we need to exploit, one of the (6) or (9) with corresponding parameters, which are related to the special case $\xi \downarrow 0$. The estimation of upper end point $\hat{\omega}$ has been studied thoroughly in [11]. Below we present the formulas to get an upper end point estimate, as well as the corresponding confidence intervals to measure the accuracy of the estimate, from the sample of maxima or the sample of exceedances for the cases $\xi \downarrow 0$ and $\xi > 0$ respectively.

For the first case $\xi > 0$, upper end point estimate can be obtained as follows:

$$\hat{\omega}(F) = \hat{\sigma_u} + u \quad (10)$$

where $\hat{\sigma_u}$ is *Maximum Likelihood* (ML) estimate of parameter σ_u that characterizes the Beta *probability density function* (pdf).

As we discussed in the previous subsection, the pdf of the Beta family (8) ($gp(x) = \frac{dGP}{dx})_{\xi>0}$) is the limiting distribution that models asymptotically the sample of exceedances over a threshold when $\xi > 0$. The corresponding log-likelihood function of a Beta-distributed sample of size k is:

$$logL(\sigma_u,\beta)=\sum_{i=1}^{k}\Big(log\frac{\beta}{\sigma_u}+(\beta-1)log\Big(-\frac{(X_i-u)-\sigma_u}{\sigma_u}\Big)\Big) \tag{11}$$

Maximization of (11) with respect to σ_u and β yields estimates $\hat{\sigma_u}$ and $\hat{\beta}$. The confidence interval that corresponds to a confidence level of $(1-\delta)*100\%$ is [12]:

$$\mid \hat{\omega}(F)-\omega(F)\mid \leq \frac{z_{\delta/2}}{\sqrt{r}}\hat{\sigma_u}(\hat{\beta}-1)\sqrt{\frac{\hat{\beta}-2}{\hat{\beta}}} \tag{12}$$

where $z_{\delta/2}$ is the $\delta/2$ quantile point of the standard normal distribution $N(0,1)$.

We follow the exceedances method instead of the maxima method when $\xi>0$ for two reasons. The first one is because the Beta distribution for exceedances in (8) is a function of two parameters, whereas the Weibull distribution for maxima in (5) is a three-parameter function and as a result the Beta log-likelihood function is more convenient to optimize. The second reason is because we expect the size of exceedances sample to be greater than the size of maxima sample and consequently to give a better quality in the final estimate.

On the other hand, the lack of a parametric expression of $\omega(F)$ in the exceedances approach renders the maxima method, for which $\omega(F)$ appears as a parameter (see σ_m in eq. (6)), our only option when $\xi\downarrow 0$. An estimate for the upper end point, when $\xi\downarrow 0$, can be given from the following formula:

$$\hat{\omega}(F)=\hat{\mu_m}+\frac{\hat{\sigma_m}}{1+\sqrt{\pi\log l}(\text{erf}(\sqrt{\log l})-1)} \tag{13}$$

where $\text{erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$ is the well-known *error function* and $\hat{\mu_m}$, $\hat{\sigma_m}$ are the ML estimates of μ_m and σ_m of Gumbel, respectively.

The pdf of the Gumbel family (6), as we discussed in the previous subsection, is the limiting distribution that models asymptotically the sample of maxima when $\xi\downarrow 0$. The corresponding log-likelihood function of a Gumbel-distributed sample of size m is:

$$\log L(\mu_m,\sigma_m)=-\sum_{i=1}^{m}(\frac{X_i-\mu_m}{\sigma_m}+\exp(-\frac{X_i-\mu_m}{\sigma_m})+\log\sigma_m) \tag{14}$$

Both $\hat{\mu_m}$ and $\hat{\sigma_m}$ can be obtained by the maximization of (14). A confidence interval for the estimate (13) (corresponding to a confidence level of $(1-\delta)*100\%$) is given by:

$$\mid \hat{\omega}-\omega\mid \leqslant \frac{z_{\delta/2}}{\sqrt{n}}\frac{\hat{\sigma_n}\sqrt{6}}{\pi}\cdot\sqrt{(\gamma-1)^2+\frac{\pi^2}{6}+\frac{2(1-\gamma)}{s_m}+\frac{1}{s_m^2}} \tag{15}$$

where $s_m=1+l\sqrt{\pi\log m}(\text{erf}(\sqrt{\log m})-1)$, $z_{\delta/2}$ is the δ/2-quantile point of the standard normal distribution and $\gamma\simeq 0.5772\ldots$ is the *Euler gamma* constant.

In order to complete the discussion about upper end point estimation, we have to show how to determine which of the two cases $\xi>0$ or $\xi\downarrow 0$ takes place, given a sample $\mathbf{X}_n$, as the parent cdf $F(x)$ is, most likely, unknown. To this end, the authors in [11] propose a test statistic $T(\mathbf{X})$ and a rejection region C_α which corresponds to a given significance level α to question the null hypothesis $H_0:\xi\downarrow 0$ against the alternative hypothesis $H_1:\xi>0$. If $T(\mathbf{X})$ falls into C_α, then H_0 is rejected at this particular significance level. Given the cdf $F_T(x)$ of $T(\mathbf{X})$ under the null hypothesis, the critical region takes the form $C_\alpha=\{\mathbf{X}:T(\mathbf{X})\geq F_T^{-1}(1-\alpha)\}$ so that the probability of $T(\mathbf{X})$ falling into C_α is equal to α. The test statistic presented below, when evaluated on the sample of exceedances (consisting of k units samples), tends asymptotically (as $k\to\infty$) to a normal distribution under the null hypothesis H_0:

$$T_{\mathbf{X}_{ex}}=\frac{\frac{(m_{\mathbf{X}_{ex}}-u)^2}{s^2_{\mathbf{X}_{ex}}}}{\frac{2}{\sqrt{k}}}\sim N(0,1) \tag{16}$$

where $m_{\mathbf{X}_{ex}}$ and $s_{\mathbf{X}_{ex}}$ are the mean and standard deviation of $\mathbf{X}_{ex}$. Accordingly, the critical region for a one-tailed test [12] is:

$$C_\alpha=\mathbf{X}_{ex}:T(\mathbf{X}_{ex})\geq z_\alpha \tag{17}$$

where z_α is as usual, the α quantile point of the standard normal distribution $\sim N(0,1)$.

III. Worst Case Delay Analysis

The worst-case condition for one stage, when studied in isolation of the rest of the circuit, may provide the best-case conditions for the previous stage. In this section we demonstrate why it is infeasible to extract a global worst-case delay condition apriori.

A. Impact of Design Parameter Variations on Maximum Delay

The delay of a path (D) can be calculated by adding all the individual delays corresponding to gates & wires on the path. More specifically, the delay of the i-th gate on a path (d_i) is proportional to the total driving capacitance C_L given by:

$$C_L=C_w+\sum_{j\epsilon fanout(i)}C_g(W_j),\ C_g\propto W \tag{18}$$

where C_w corresponds to the total wire capacitance seen by the gate and C_g corresponds to the input pin capacitances of the fanout gates, and it is inversely proportional to the drain-source I_{ds} current of the driving transistors ($I_{ds}\propto(W,1/L,f(V_{th}))$, where $f(V_{th})$ is a strictly decreasing function for the V_{th} range of interest) [13].

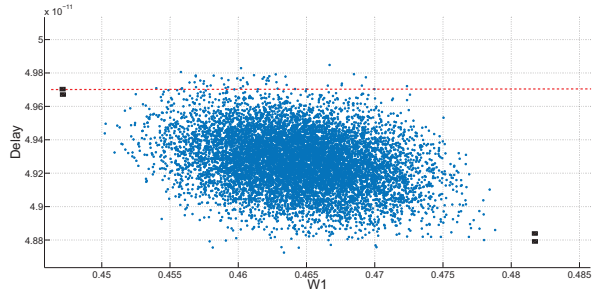
Under process variation, parameters that characterize the electrical behavior (W,L,V_{th}) of the transistor deviate from the nominal values, in some cases substantially, thus we have to treat them as random variables distributed between a lower and an upper bound, when analyzing circuit performance. A sound assumption, made in other similar studies, about the distribution of W, L and V_{th} is that are normally distributed between $\pm 3\sigma$. So we consider $W\epsilon[W_0-3*\sigma_w,W_0+3*\sigma_w]$, $L\epsilon[L_0-3*\sigma_L,L_0+3*\sigma_L]$ and $V_{th}\epsilon[V_{th0}-3*\sigma_{Vth},V_{th0}+3*\sigma_{Vth}]$ as normal random variables with mean values W_0, L_0 and V_{th0} and standard deviation σ_w, σ_L and σ_{Vth} respectively.

Therefore, when the effective width of a gate on a path is increased, the driving current is increased and as a result its delay is decreased. On the other hand, the delay of the previous gate is increased, because it bears a higher load. Also, σ_{Vth} is inversely proportional to the square root of W and L ($\propto 1/\sqrt{WL}$), according to Pelgrom's rule [14]. So when L is increased, $\sigma_{V_{th}}$ and I_{ds} are decreased, which indicate that the delay of the gate increases and at the same time the worst case delay condition coming from V_{th} ($V_{th0}+3*\sigma_{Vth}$) becomes smaller.

Note that similar arguments hold for the interconnects. Interconnect capacitance and resistance are a function of parameters that are determined by the interconnect structure such as width, thickness and spacing, which in turn can be

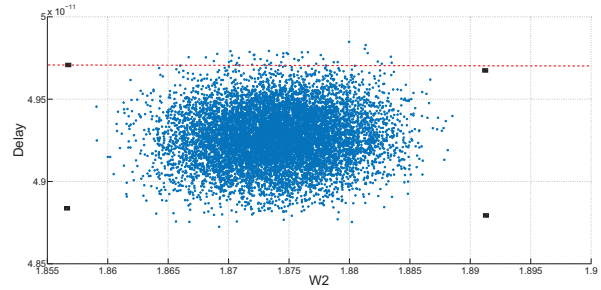
considered as bounded random variables normally distributed. For example, figure 3 in [5] highlights that the condition of the maximum interconnect capacitance is opposite to that for maximum resistance and RC constant, an observation that is aligned to the previous discussion about transistor parameters.

Figure 1 shows the maximum delay of two inverters in a row, taken from the memory decoder assumed by Cacti [15], where $\sigma_{w_1} = 5\%$ and $\sigma_{w_2} = 5\%$ of minimum transistor size. Notice that the maximum delay is somewhere between the corners of $\mathbf{W} = \{W_1, W_2\}$, and thus, cannot be determined in advance. Even if the worst delay was located at the corners of W space, the number of all possible process parameters combinations grows exponentially with the circuit size.



(a) The projection of a 3d scatter plot of two-inverters delays on W1 axis.

.



(b) The projection of a 3d scatter plot of two-inverters delays on W2 axis.

Fig. 1: The figure is a scatter plot of two-inverters delay derived from a MC simulator with 10,000 different (W1, W2) pairs. We denote the delays at W1, W2 corners with black squares. Notice that there are a lot of points over the red-dotted line, which indicates the maximum delay from the set of delays corresponding to $(W1, W2)$ corners, namely $\{(W1_{min}, W2_{min}), (W1_{min}, W2_{max}), (W1_{max}, W2_{min}), (W1_{max}, W2_{max})\}$.

IV. Proposed Methodology for Maximum Delay Estimation

Following the discussion in the previous section, the problem focuses on finding the maximum (upper end point) of a bounded random variable on the basis of a random sample. This problem has been tackled successfully in [11] (for the problem of maximum current estimation). In this section we review the previous methodology and comment on how it can be applied on the problem of maximum delay estimation. In addition, we share some initial thoughts for further enhancements of the previous methodology.

The first step of the methodology entails a MC simulation of n trials to acquire the statistical sample of ATs at the POs $\mathbf{AT}_n = \{AT_1, AT_2, \ldots, AT_n\}$, where each MC trial is performed with a random $\mathbf{P} = \{P_1, P_2, \ldots, P_{\#dp}\}$ sample (P_i is assigned to the i_{th} design parameter, which in turn belongs to either a gate or interconnect) drawn from the joint probability $F_{\mathbf{P}} \sim (P_1, P_2, \ldots, P_{\#dp})$. The second step is to sort the units of the AT sample in ascending order and pick those unit samples that exceed a threshold u (makes the sample of exceedances $\mathbf{X}_{ex}$). The threshold u has to be selected so that the sample of exceedances reside in the tail of AT distribution. Therefore, the units of $\mathbf{X}_{ex}$ have to be as many as the parameters of GP pdf, which is the limiting distribution of the sample of exceedances, can be approximated perfectly. On the other hand, when a too small threshold u is selected in order to include more sample units of the initial AT sample, then there is the peril including in X_{ex} units that do not belong in the tail of AT distribution. A rule of thumb is to set the threshold u so that the upper 10% of the initial AT sample is included to $\mathbf{X}_{ex}$. More rigorous procedures for automatic selection of u can be found in [16], [17]. Based on the sample $\mathbf{X}_{ex}$ we calculate the test statistic T in (16) and compare it with the critical value of z_α (for significance level α) to decide whether we accept or reject the hypothesis $\xi \downarrow$. Experiments performed on various random samples drawn from an Exponential distribution have resulted to a critical value of $z_\alpha \simeq 7$ which corresponds to significance level of $\alpha = 10^{-12}$ (detailed proof of the previous result can be found in Appendix C of [11]). Subsequently, if the hypothesis $\xi \downarrow$ is true we divide the initial sample of AT into l blocks of size m sub-samples and pick the the maximum unit from each sub-sample forming the sample of maximas. Then, we calculate the ML estimates $\hat{\sigma}$ and $\hat{\mu}$ of the corresponding Gumbel parameters σ, μ as the solution given from the maximization of (14) over the sample of maximas and finally determine the maximum AT for the previous estimates and provide the confidence interval for the chosen confidence level $1-\delta$ from (13) and (15) respectively. On the other hand, if the results of test statistic T indicate that we have to reject the hypothesis $\xi \downarrow$ ($T \geq 7$), we find the ML estimates $\hat{\sigma}$ and $\hat{\beta}$ of the corresponding Beta parameters σ, β as the solution given from the maximization of (11) over the sample of exceedances and finally we determine the maximum AT for the previous estimates and provide the confidence interval for the chosen confidence level $1-\delta$ from (10) and (12) respectively. Below we summarize the steps of the methodology for maximum AT estimation we described previously:

Step 1: Generate n **P** samples and assign them to the corresponding gates or interconnects.
Step 2: Perform a MC simulation with n trials to acquire the statistical sample of ATs at the POs $\mathbf{AT}_n = \{AT_1, AT_2, \ldots, AT_n\}$.
Step 3: Sort the units of AT sample in ascending order and pick the units that are over a threshold u (sample of exceedances).
Step 4: Evaluate the test statistic T in (16) for the sample of exceedance acquired in the previous step.
Step 5: if ($T < 7$):
  Step 5.a: Derive the sample of maximas from the sample of ATs.
  Step 5.b: Estimate the Gumbel parameters by maximizing (14) over the sample of maximas.
  Step 5.c: Determine the maximum AT estimate from (13) and the confidence interval for the chosen confidence level $1-\delta$ from (15).
Step 6: if ($T \geq 7$):
  Step 6.a: Estimate the Beta parameters by maximizing (14) over the sample of exceedances.
  Step 6.b: Determine the maximum AT estimate from (10) and the confidence interval for the chosen confidence level $1-\delta$ from (12).

A. Enhancements to the Proposed Methodology

The previous methodology can be enhanced with machine learning techniques to determine the size n of MC simulations, which dominantly affects the time complexity of the methodology, as well as the threshold u. For example we can acquire the sample of exceedances at step 3 exploiting the Statistical Blockade algorithm presented in [18]. Statistical Blockade builds a classifier based on an initial input variables training set of size $n_0 << n$ (in our methodology the **P** sample) and a sample of the corresponding metric of interest (in our case the AT sample), derived after the simulation of the underlying system (in our case SSTA), to test whether a unit sample of the metric we are interested in, would fall in the tail of the parent distribution for every new unit sample of input variables (**P**), beyond the first n_0 samples, before doing a simulation. By doing so, for a good classifier, we acquire a sample of exceedances that falls in the tail of the parent distribution with greater probability than simply selecting the upper 10% of the initial AT sample, thus speeding up the methodology and improving the accuracy simultaneously. Notice that if $\xi \downarrow$ is the case, then we cannot rely on the sample of exceedances. For this reason we keep a track of the rejected **P**s by the Statistical Blockade algorithm, run the missing simulations and follow subsequently the rest of the steps corresponding to case $\xi \downarrow$. Again, because we will have rejected many more P samples than those leading to AT samples in the tail, we will have both greater accuracy and speedup, as n becomes a function of u.

V. Experimental Results

For the experimental evaluation we developed an SSTA tool in C++, based on MC simulations, to gather the statistical sample of ATs at different POs. Our SSTA tool embeds models that allow us to variate only the effective transistor width of a gate (W). However, that does not prevent us from deriving sound conclusions about the effectiveness of the proposed methodology, as the methodology does not depend on the number of design parameters that affect timing but on the statistical sample at the outputs. For timing verification where the worst slack is required, the extension is trivial by substituting the ATs sample in the proposed methodology with the SLACKs sample, as the i.i.d assumption we make for the sample is not violated. We applied our methodology on a subset of ISCAS85/89 [19][20] benchmarks implemented at a 90nm technology node. For the evaluation we used a 3.60GHz Intel Core i7-4790 CPU with 16 GB memory system running UNIX. Below we summarize the steps we followed to run the experiments.

- We run SSTA with n trials to obtain the statistical sample of ATs at each PO ($\mathbf{AT}_k = \{AT_{k,1}, AT_{k,2}, \ldots, AT_{k,n}\}$, where k denotes the k-th PO) of the circuit under test.
- We apply steps from 3 to 6 of our methodology on $\mathbf{AT}_k$ to get an estimate of max AT_k ($\hat{AT_k}$) and the corresponding confidence interval $\mid \hat{AT_k} - AT_k \mid$.
- Finally, we evaluate the current relative error $\left(\eta = \frac{\mid \hat{AT_k} - AT_k \mid}{\hat{AT_k}}\right)$ of the the worst AT_k estimate.
- If η is below a predefined target relative error ($\eta_{target} = 5\%$), we acquire more samples from the SSTA tool and re-apply our methodology until the desired accuracy is achieved.

The generation of $\mathbf{W}_{\#gates}$ samples in Step 1 of our methodology can be done very fast by randomly picking a corner W_k for each gate. The accuracy of the estimated ATs in Step 2 is entirely up to the timing analysis tool and varies from SPICE to gate-level. Also, the computational time required to complete Step 2 depends on the efficiency of the timing engine employed. However, the runtime of Steps 3-6 is very small compared to the runtime of MC simulation, and thus, one can re-arrange each sample $\mathbf{AT_k}$ into sub-samples of various sizes in order to obtain better estimates.

The results from the execution of the above steps, in the case of three sequential (s27, s35932, s38417) and three combinational (c17, c6288, c7552) designs, are reported below. In this experiment, each MC trial is performed with a random $\mathbf{W}_{\#gates} = \{W_1, W_2, \ldots, W_{\#gates}\}$ sample, where each W is normally distributed between $\pm 3\sigma$. Table I shows the maximum of **AT** sample and the estimated $\hat{AT}$ at selected POs, for each circuit under consideration, when the target relative estimation error is within 5% for 99% confidence level. Additionally, it reports the AT sample size and the achieved relative estimation error η.

It is worth pointing out that the random samples required to meet this relative estimation error for POs that belong on paths of higher impedance interconnects are slightly increased. Such circuits include clock network interconnects that are rooted on the upper level metal layers where the resistance is lower. Resulting sample sizes for the above benchmarks, implemented using both high and low impedance interconnects, are presented in Table II. In the former case a SPEF file is extracted, while in the latter case a zero wire delay model is assumed. Note that in both cases estimated AT at the POs are within 5% of the true maximum AT at a cost of a few thousand MC simulations.

TABLE I: Required sample/sub-sample sizes, sample and estimated maximum AT on a subset of POs when relative estimation error is within 5% for 99% confidence level.

Circuit	Primary Output	Sample Size	Sub-Sample Size	Sample max (ns)	Estim. max (ns)	Relative Error (%)
s27	G17	2500	25	0.265	0.270	0.47
s35932	DATA_9_0	5000	25	0.513	0.694	4.25
	DATA_9_19	5000	25	0.518	0.702	4.26
s38417	g16297	2500	50	0.132	0.146	2.99
	g25420	10000	25	0.470	0.820	0.75
c17	N22	2500	25	0.053	0.054	0.44
	N23	2500	25	0.051	0.052	0.07
c6288	N5971	10000	25	2.304	2.691	1.62
	N6280	5000	25	3.428	3.940	2.11
c7552	N10718	5000	25	0.917	0.940	0.42
	N10729	5000	25	0.671	0.809	2.73

TABLE II: Sample size required in order to achieve 5% relative estimation error for 99% confidence level. Higher impedance interconnects require more MC simulations.

Circuit	Sample Size (with interconnect	Sample Size (w/o interconnect)
s27	2500	2500
s35932	5000	2500
s38417	10000	2500
c17	2500	2500
c6288	10000	5000
c7552	5000	5000

To evaluate the efficiency of our statistical method, we compare it against an exhaustive MC simulation, which corresponds to the slowest but full accurate version of corner-based analysis (a.k.a. full factorial design), for a test benchmark. On the exhaustive approach we have to simulate all the combinations of gate widths, while when we apply our method we choose n random samples of them. If we allow W to take a value within a continuous range, then the number of all combinations is not bounded. For this reason, each gate width W is set on its $\{-3\sigma, +3\sigma\}$ corner. As a result, the max AT derived from the exhaustive MC simulation is a lower bound of the actual max AT, and thus, the reported relative error is an upper bound of the actual one.

In order to perform this experiment, we pick a design with 30 gates where the exhaustive MC can be feasible since even for small designs of 100 gates the runtime would be tremendous (2^{100} MC trials). Due to the lack of availability

of such a well-known circuit, we implemented a synthetic benchmark, based on the ISCAS c432 benchmark, which consists of 30 gates.

We first demonstrate the accuracy of our methodology and report the relative estimation error on a single selected PO (N223) of the test design. In Table III, we present how our methodology approaches the theoretical max AT at N223. From the table, we can clearly observe that the more samples we pick to apply our method on, the better estimation and relative error is achieved.

Note that the reported relative error stands for the upper bound of the computed confidence interval, thus for a worst-case timing analysis we have to add this error to the estimated maximum AT at the PO.

TABLE III: Relative error on N223 for different sample/sub-sample sizes and confidence level 99%.

Sample Size	Sub-Sample size	Sample max (ns)	EVT max (ns)	Relative Error (%)
2500	50	0.16784	0.16997	0.540
5000	25	0.16833	0.16987	0.289
10000	50	0.16811	0.16956	0.235
20000	50	0.16814	0.16818	0.049
50000	25	0.16814	0.16972	0.088
100K	25	0.16848	0.16971	0.062
1M	50	0.16848	0.16890	0.016
10M	50	0.16848	0.16953	0.007
100M	25	0.16848	0.17068	0.002

Figure 2 depicts the probability distribution at node N223 after 10 million MC simulations. The following distribution is a representative one of all POs in all benchmarks we tested and corresponds to the case $\xi \downarrow$ in the proposed methodology. Also, following the previous observation, another point to mention is that in all experiments we find that a number of 50 sub-samples, in block maxima modelling, yields estimates with relative error of about 5% (at a confidence level of 99%) for any PO irrespective of its level.

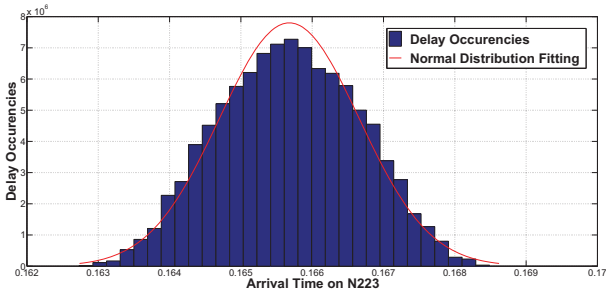


Fig. 2: Probability Distribution of the AT on N223 after MC simulation of 10M trials.

The runtime efficiency of the proposed method is reported in table IV . For this comparison we execute the previous experiment for N223 and measure the runtime.We conclude that compared to an exhaustive MC simulation, our method achieves up to six orders of magnitude (x*233426*) better performance. In the case targeting a 5% relative estimation error for all POs, given that experimental results on ISCAS show that we need a maximum of 10000 samples (Table II), the proposed method can be about five orders of magnitude (x*58355*) faster than an exhaustive MC.

VI. Conclusion

In this paper we present a novel statistical methodology based on EVT, as a substitute of the conventional corner-based analysis, which can provide accurate estimates of the worst

TABLE IV: Runtime comparison between the proposed methodology and an exhaustive MC simulation.

Sample Size	Our Method Runtime (sec)	Exhaustive MC Runtime (sec)	Speedup
2500	0.062	14472	x233420
5000	0.124	14472	x116710
10000	0.248	14472	x58355
20000	0.496	14472	x29177
50000	1.239	14472	x11680
100K	2.479	14472	x5838
1M	24.898	14472	x582
10M	248.527	14472	x59
100M	2478.140	14472	x5.85

ATs at the POs taking into account process variation. Our methodology does not make any assumptions about the gate or interconnect timing model or the distribution of AT at the POs, and thus, can be used from gate to transistor level of abstraction in timing verification. Experimental results showed that we can achieve a relative error between the true maximum and the estimated maximum that is below 5% with just a few Monte Carlo simulations.

Acknowledgment

The authors would like to thank Michalis Tsiampas for his fruitful comments on this work.

References

[1] H. Terada et al., *Accurate Estimation of the Worst-Case Delay in Statistical Static Timing Analysis*, IPSJ, Transactions on Systems LSI Design Methodology, Vol.1, pages 116-125, August 2008.
[2] Abhijit Dharchoudhury and S. M. Kang, *Worst-case Analysis and Optimization of VLSI Circuit Performances*, IEEE Transactions On Computer-Aided Design Of Integrated Circuits And Systems, Vol. 14, No. 4, April 1995.
[3] Hong Zhang et al., *Efficient Design-Specific Worst-Case Corner Extraction for Integrated Circuits*, DAC 2009, July 26-31.
[4] Xiang Lu et al., *PARADE: PARAmetric Delay Evaluation Under Process Variation*, International Symposium on Signals, Circuits and Systems. Proceedings, SCS 2003.
[5] Takayuki Fukuoka et al., *Worst-case Delay Analysis Considering the Variability of Transistors and Interconnects*, Proceedings of the 2007 International Symposium on Physical design, Pages 35-42.
[6] R.-D. Reiss and M. Thomas, *Statistical Analysis of Extreme Values*, Boston, MA:Birkhauser.
[7] Fisher RA, Tippett LHC *Limiting forms of the frequency distribution of the largest or smallest member of a sample*, Proceedings of the Cambridge Philosophical Society, 24:180190, 1928.
[8] Balkema AA, de Haan L, *Residual life time at great age*, The Annals of Probability, 1974.
[9] Pickands J III AA, de Haan L, *Statistical inference using extreme order statistics*, Annals of Statistics, 1975.
[10] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*, 2nd ed: Krieger, 1987.
[11] Nestoras E. Evmorfopoulos et. al., *A Monte Carlo Approach for Maximum Power Estimation Based on Extreme Value Theory*, IEEE Transactions on Computer-Aided Design of Integrated Circuits, vol 21, April 2002.
[12] G. Rousas, *A Course in Mathematical Statistics*, 2nd ed. New York: Academic, 1997.
[13] Neil Weste and David Harris, *CMOS VLSI Design: A Circuits and Systems Perspective*, Addison-Wesley Publishing Company.
[14] M. Pelgrom et al., *Matching properties of mos transistors*, Solid-State Circuits, IEEE Journal of, vol. 24, no. 5, pp. 14331439, Oct 1989.
[15] Steven J. E. Wilton and others, *CACTI: An Enhanced Cache Access and Cycle Time Model*, IEEE JSSC, 1996, vol. 31, pp. 677-688.
[16] H. Drees and E. Kaufmann Zhang et al., *Selecting the optimal sample fraction in univariate extreme value estimation*, Stochastic Processes and their Applications, vol. 75, pp. 149-172, 1998.
[17] A. Guillou and P. Hall, *A diagnostic for selecting the threshold in extreme value analysis*, J. Royal Statist. Society B, vol. 63, pp. 293-305, 2001.
[18] A. Singhee et al., *Statistical blockade: very fast statistical simulation and modeling of rare circuit events and its application to memory design*, IEEE Trans. Computer-Aided Design, vol. 28, no. 8, pp. 1176-1189, 2011.
[19] F. Brglez and H. Fujiwara, *A Neutral Netlist of 10 Combinational Benchmark Circuits and a Target Translation in Fortran*, IEEE International Symposium on Circuits and Systems, pp. 1929-1934, 1985.
[20] F. Brglez et al., *Combinational profiles of sequential benchmark circuits*, IEEE International Symposium on Circuits and Systems, vol. 3, pp. 1929-1934, 1989.