Circuit Propagation Delay Estimation Through Multivariate Regression-Based Modeling Under Spatio-Temporal Variability

Shrikanth Ganapathy[†] Ramon Canal[†] Antonio Gonzalez[†‡] Antonio Rubio[§]
[†]Department d'Arquitectura de Computadors Universitat Politècnica de Catalunya {sg,rcanal}@ac.upc.edu
[‡]Intel Barcelona Research Center Intel Labs-UPC antonio.gonzalez@intel.com
[§]Department d'Enginyeria Electrònica Universitat Politècnica de Catalunya antonio.rubio@upc.edu

***Abstract*—With every process generation, the problem of variability in physical parameters and environmental conditions poses a great challenge to the design of fast and reliable circuits. Propagation delays which decide circuit performance are likely to suffer the most from this phenomena. While Statistical static timing analysis (SSTA) is used extensively for this purpose, it does not account for dynamic conditions during operation. In this paper, we present a multivariate regression based technique that computes the propagation delay of circuits subject to manufacturing process variations in the presence of temporal variations like temperature. It can be used to predict the dynamic behavior of circuits under changing operating conditions. The median error between the proposed model and circuit-level simulations is below 5%. With this model, we ran a study of the effect of temperature on access time delays for 500 cache samples. The study was run in 0.557 seconds, compared to the $20h\ and\ 4min$ of the SPICE simulation achieving a speedup of over $1X10^5$. As a case study, we show that the access times of caches can vary as much as $2.03X$ at high temperatures in future technologies under process variations.**

I. Introduction

In the last 40 years, microprocessor design has obediently followed moore's law by allowing the integration of twice the number of transistors every eighteen months. Moore's law mainly deals with scaling of resources quantitatively and does not take into account the changing device characteristics. With reducing feature sizes, while it is possible to follow the moore's law faithfully by building more transistors on the same chip area, transistors themselves do not behave faithfully [1]. This is mainly due to the difference in physical parameter values of the transistors between the intended design and obtained chip. This deviation termed as process variations, is increasingly important and a few generations from now, it will be the most important design challenge.

According to the recent ITRS reports [2], variability has been identified as a key design challenge in the coming years. A thorough understanding of manifestation of such variations is imperative and their impact on the performance of circuits has to be studied in order to scale the performance of circuits in future. Not coping up with such challenges will only mitigate the advantages of device scaling.

In order to understand the impact of variability on circuit performance, we need to see its effect on the two most important variables - power and delay [3]. While power has been a major concern for a very long time from an energy efficiency point of view, there is very little knowledge about the effect on delay due to combined effects of spatio -and- temporal variations. Also, most often when researchers refer to *parameter variations*, they take into account only the changes in physical parameters like threshold voltage (V_{th}) and effective channel length (L_{eff}) and do not to account for changing operating conditions like temperature and inputs. In the context of parameter variations, spatial variations refer to the changes in physical parameter values across the die and temporal variations refer to the changing operating conditions of the die.

Spatial variations are not only required for process optimization and control but also help in designing better circuits that are robust to these variations [4]. They arise due to non-ideal manufacturing process [5], [6], [7]. On the other hand, temporal variations can occur at a frequency of nanoseconds to years [3]. The phenomena is furthered by circuit performance and operating conditions. We also know that variation in one domain can cause a variation in the other domain [8]. Traditionally, the method adopted towards understanding the manifestation of spatio-temporal (ST) variations is to manufacture test chips and subject them to rigorous testing under enhanced operating conditions. The method also called Burn-in [9], accelerates the production of faults and exposes them at test time. In this method, it is hard to extrapolate the behavior of circuits subjected to ST variations for future technologies.

In this paper, we propose a novel technique to exploit multivariate regression analysis for predicting the deviations in propagation delay due to combined effects of spatio-temporal variations. We evaluate the error between the results obtained from our model and results from circuit level simulations. Among other applications, the model may speed-up the forming of heuristics for speed binning and timing yield prediction while enabling fast and accurate design space exploration.

The rest of paper is organized as follows. In Section 2, we discuss about the related work. In Section 3, we propose an extension to the existing multi-level scheme for modeling of spatial variations [10] and discuss about temporal variations in brief. In Section 4, we discuss the multivariate regression based modeling technique to compute the delay of circuits under spatio-temporal variations. In Section 5, we present the evaluation methodology for testing the regression model on a 32KB cache. In section 6, we then validate our technique by computing the error in calculation of propagation delay between SPICE simulations and the regression model. In section 7, we present 2 case studies exploring the wide-usage of the proposed model. Finally, in Section 8, we present the concluding remarks.

II. Related Work

Currently, propagation delay characterized by SSTA techniques are modeled as a Gaussian function dependent on physical parameter variation [11], [12], [13]. However, delay is very much dependent on circuit temperature and gate inputs. These works assume that gates have similar parameter distribution failing to account for random variations across adjacent transistors within a gate. In [14], a graph-based methodology is employed and it captures every source of variation for computing the final arrival times and also the sensitivity to each of the sources of variation. The methodology neglects the independent effects of temporal variations. In [15], the effect due temperature variation on delay is examined as a function of decreasing supply-threshold ratio.

III. Modeling of Spatio-Temporal Variations

A. Modeling of Spatial Variations

Spatial variations can be assumed to be the summation of Die-to-Die(D2D), Within-Die(WID) Systematic and Within-Die Random

components. D2D variations are known to affect devices within the same die similarly. Systematic variations are deterministic in nature. Random variations on the other hand are caused due to dopant fluctuations and line edge roughness [3]. Their effect is modeled by assuming a finite amount of parameter variation which is Gaussian distributed. Sarangi et al. [16] have modeled the correlation in parameter values between 2 points as a cubic function of the distance between them. It is an analytical model and the results have been validated against empirical data. Agarwal et al. [10] have proposed a Multi-Level QUADTREE based modeling technique that reverse engineers the empirical residual modeling proposed by Stine et al. [4]. We begin with [10] as the base for our model and extend it based

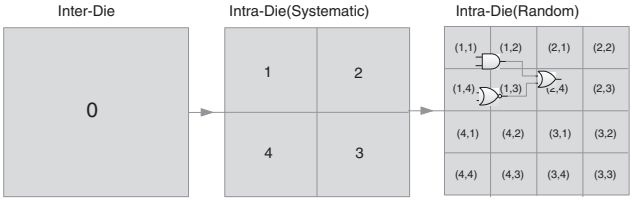

Fig. 1. 2D View of the QUADTREE Implementation.

on the following assumptions:
1) Correlation between spatial points is a function of similarity of structures also. Structures having similar layout geometries will be affected in a similar fashion [13]. For e.g. columns of replicated SRAM cells will have similar parameter distribution.
2) Physical parameters taken into account are V_{th} and L_{eff} are assumed to be Gaussian distributed. Variation in other parameters can be assumed as a function of the variation of the above parameters.
3) Worst case deviation of parameter values is rare. By virtue of the additive nature of the model, realistic predictions about the distribution of parameter values can be made.
As show in Figure 1, every grid $G_{i,j}$ in the last level is linked to grid G_i in the second level and every grid in this level is linked to G_0. The value of a parameter $P_{i,j}$ is given by:

$$\begin{aligned} P_{i,j} &= P_{nominal} + \delta P_{variation} \\ &= P_{nominal} + \delta P_{inter-die} + \delta P_{intra-die} \\ &= P_{nominal} + \delta P_{inter-die} + \delta P_{systematic_i} \\ &+ \delta P_{random_{i,j}} \end{aligned} \quad (1)$$

Similar layout structures are grouped together in the same grid present in the level above the last level. All points within a grid are assumed to have a similar parameter distribution. This way, distance based correlation is achieved by virtue of the model and layout based correlation by virtue of the specification of the model. The simulation framework is shown in Figure 2.

B. Motivation to Model Temporal Variations

Thermal management plays a vital role in the design of circuits [17]. Sudden temperature shoot-ups can result in functionality and reliability problems. Spatially, temperature varies from one region to another depending on hotspot generations. Temporally, they vary depending on the computing workload and dissipated power. Although enough guard banding is provided to accommodate for worst case temperatures, with every process generation, a generation of performance is being lost due to ST variations [3]. In order to validate our assumption that variations in temperature would affect circuit performance to a very great extent, we designed a 32KB cache and measured the access times across five generations for different

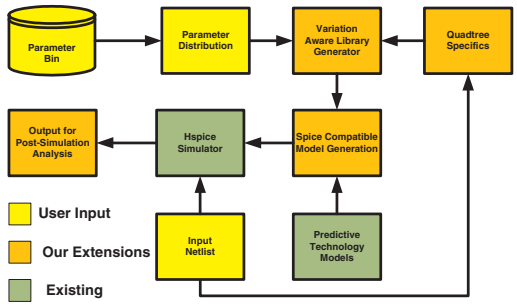

Fig. 2. HSPICE Simulation Framework

temperatures with Predictive Technology Models [18] as shown in the Figure 3. The simulation is performed only for temporal variations of temperature and spatial variations of process parameters are not taken into account. It can be seen that, with increase in temperature the difference in access time widens for every process generation. This behavior is due to the decreasing difference in values of V_{dd} and V_{th} with every process generation that results in an overall increase in delay. The variation of V_{th} due to temperature variations as modeled

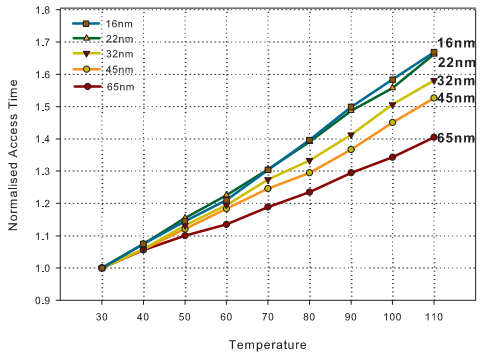

Fig. 3. Access Time variation for Temperature Variation

by BPTM is given by

$$V_{th(temp)} = V_{th(t_{ref})} + (KT1 + KT2.V_{bs}).(T_{ratio} - 1) \quad (2)$$

In the above Equation 2, KT1 and KT2 are model parameters, temperature and body bias coefficient respectively. T_{ratio} is dependent on temperature and increases with increase in temperature. From the Equation 2 it is clear that: as temperature rises, threshold voltage degrades resulting in reducing supply-threshold ratio making circuits slower. Thus it is extremely important to understand the interactions and ramifications of this vicious circle.

IV. Path-Aware Delay Modeling

A. Transistor Specific Delay Modeling

Commercial CAD tools perform timing analysis of gates using lookup tables built during library characterization. For every combination of input slew and output loading, the delay is obtained as a function of the gate length [13]. The methodology as shown in Figure 4(a) assumes that all transistors within the gate have similar L_{eff} and V_{th} values. This assumption leads to a single distribution for delay as a function of varying parameters. Also, the impact of delay due to temperature variations on delay is higher when compared to other physical parameters. In reality, what happens is that gates have multiple distributions depending on their current state as shown in Figure 4(b). With increase in temperature, the probability density

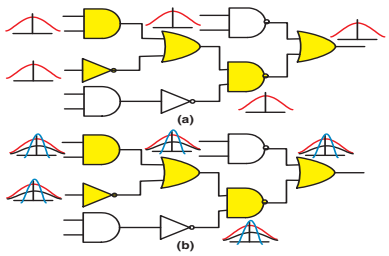

Fig. 4. Path Based Delay Calculation

function broadens and the number of samples tending towards the mean decreases with more number of samples towards right hand side of the mean.

Consider the case of 2 NOT gates connected in series. For any input, both pull-up and pull-down network of adjacent gates would actively take part in the output transition. Then again, this model would only assume a variation in the inputs and it will not account for individual variations across the transistors that constitute the path between the input and output. If we are able to determine a path for every possible input and model the parameters of every transistor along the path as a random Gaussian function, then this would result in an ideal delay model that is aware of the characteristics of every transistor that makes up the path. Nevertheless, the complexity of generating such a model would be very high. Considering the example of 2 NOT gates in series for the case where the input is 1, we extend the input based gate delay model of [11] to a path aware propagation-delay model using Equation 4,

$$\begin{aligned} D &= D_{P1} + D_{N2} \\ &= (D^{pmos}_{nominal} + \delta D_{ST_{P1}}) \\ &+ (D^{nmos}_{nominal} + \delta D_{ST_{N2}}) \qquad (3) \\ \delta D_{ST_{P1}} &= m_{leff} * \delta^{P1}_{leff} + m_{vth} * \delta^{P1}_{vth} \\ &+ m_{temp} * \delta^{P1}_{temp} \qquad (4) \\ \delta D_{ST_{N2}} &= m_{leff} * \delta^{N2}_{leff} + m_{vth} * \delta^{N2}_{vth} \\ &+ m_{temp} * \delta^{N2}_{temp} \qquad (5) \end{aligned}$$

where δD_{ST} is the variation in delay due to ST variations. Here, δD_{ST} is composed of variations due to spatial variations of V_{th}, L_{eff} and temporal variations of Temperature. Assuming a linear dependence between spatial parameters and delay [12], we can derive a more generalized equation for any path *i* composed of *j* transistors given by the Equation 6[1]

$$\begin{aligned} D_i &= \sum_{a=1}^{j}(D^{pmos}_{nominal} + m^{pmos}_{leff} * \delta^{p(a)}_{leff} + m^{pmos}_{vth} * \delta^{p(a)}_{vth} \\ &+ m^{pmos}_{temp} * \delta^{p(a)}_{temp}) + \sum_{a=1}^{j}(D^{nmos}_{nominal} + m^{nmos}_{leff} * \delta^{n(a)}_{leff} \\ &+ m^{nmos}_{vth} * \delta^{n}_{vth}(a) + m^{pmos}_{temp} * \delta^{n(a)}_{temp}) \qquad (6) \end{aligned}$$

Equation 6 can be modified to Equation 7 for same temperature devices and a nominal delay determined at design stage,

$$\begin{aligned} D_i &= \sum_{a=1}^{j}(m^{pmos}_{leff} * \delta^{p(a)}_{leff} + m^{pmos}_{vth} * \delta^{p(a)}_{vth}) \\ &+ \sum_{a=1}^{j}(m^{nmos}_{leff} * \delta^{n(a)}_{leff} + m^{nmos}_{vth} * \delta^{n(a)}_{vth}) \\ &+ j * (D^{pmos}_{nominal} + D^{nmos}_{nominal}) + m_{temp} * \delta_{temp} \qquad (7) \end{aligned}$$

$$m_{temp} = j * (\delta^{p(a)}_{temp} + \delta^{n(a)}_{temp}) \qquad (8)$$

Solving Equation 7 for solution of the form Y=Xb+ϵ and estimating b=YX^{-1} we get,

$$\begin{aligned} D_i &= \begin{vmatrix} \delta^{p(1)}_{leff} & \delta^{p(1)}_{vth} & \delta^{n(1)}_{leff} & \delta^{n(1)}_{vth} \\ \delta^{p(2)}_{leff} & \delta^{p(2)}_{vth} & \delta^{n(2)}_{leff} & \delta^{n(2)}_{vth} \\ \vdots & \vdots & \vdots & \vdots \\ \delta^{p(j)}_{leff} & \delta^{p(j)}_{vth} & \delta^{n(j)}_{leff} & \delta^{n(j)}_{vth} \end{vmatrix} \times \begin{vmatrix} m^{pmos}_{leff} \\ m^{pmos}_{vth} \\ m^{nmos}_{leff} \\ m^{nmos}_{vth} \end{vmatrix} \\ &+ j * (D^{pmos}_{nominal} + D^{nmos}_{nominal}) + m_{temp} * \delta_{temp} \qquad (9) \end{aligned}$$

Solving Equation 9 would result in the slope values that closely follow the delay distribution. As in [13], we assume in the remaining of the paper the analysis of one path (the other paths are assumed to behave similarly). In other words, the results presented show only one path derived from a given set of inputs; which is the case for regular structures such as memory structures; as well as for any balanced (i.e. similar delay paths for different inputs) design.

V. Evaluation Methodology

Considering that we have adopted Quadtree model for specification of parameter distribution, it is preferable that the circuit simulated be fairly large and sensitive to ST variations. We evaluate the model on a cache constructed from minimum geometry transistors operating at scaled supply voltages that will suffer the most the from ST variations[19]. The model would also guide the design space exploration of caches under technological constraints.

A. Cache Design

A cache as shown in Figure 5 was designed and implemented in HSPICE. It is a 32KB cache block that takes advantage of array sub-blocking to reduce the impact of variations [20]. Every sub-block is decoded using a array sub-block decoder and a row decoder decodes the row within the decoded sub-block. The global and local controller generate timing signals for the following: address generation, precharging and read/write enable. Wires are more resilient to variations when compared to logic and memory structures and hence we assume it to be a fixed amount of the obtained access time. The main idea of such an experiment is to determine at run time the deviation in access times of such a cache subjected to ST variations. As for access time calculation, we have adopted a model similar to the one implemented in CACTI [21].

$$\begin{aligned} T_{write} &= T_{control} + T_{arraysub-blockdecoder} \\ &+ T_{addressdecoder} + T_{worlinedriver} \\ &+ T_{bitlinedriver} \qquad (10) \end{aligned}$$

$$\begin{aligned} T_{read} &= T_{control} + T_{arraysub-blockdecoder} \\ &+ T_{addressdecoder} + T_{wordlinedriver} \\ &+ T_{sramtransfer} + T_{senseamplifier} \\ &+ +T_{columnmultiplexer} \qquad (11) \end{aligned}$$

It is a state of the art tool to determine the absolute access time of caches based on size, technology and supply voltage characteristics. While it is capable of calculating access times for any given temperature, it does not take into consideration the interactions of spatial

[1] m^z_y represents the slope of parameter y for device of type z
$\delta^{p(z)}_y$ represents the delay deviation due to parameter y for the z^{th} pmos device
$\delta^{n(z)}_y$ represents the delay deviation due to parameter y for the z^{th} nmos device

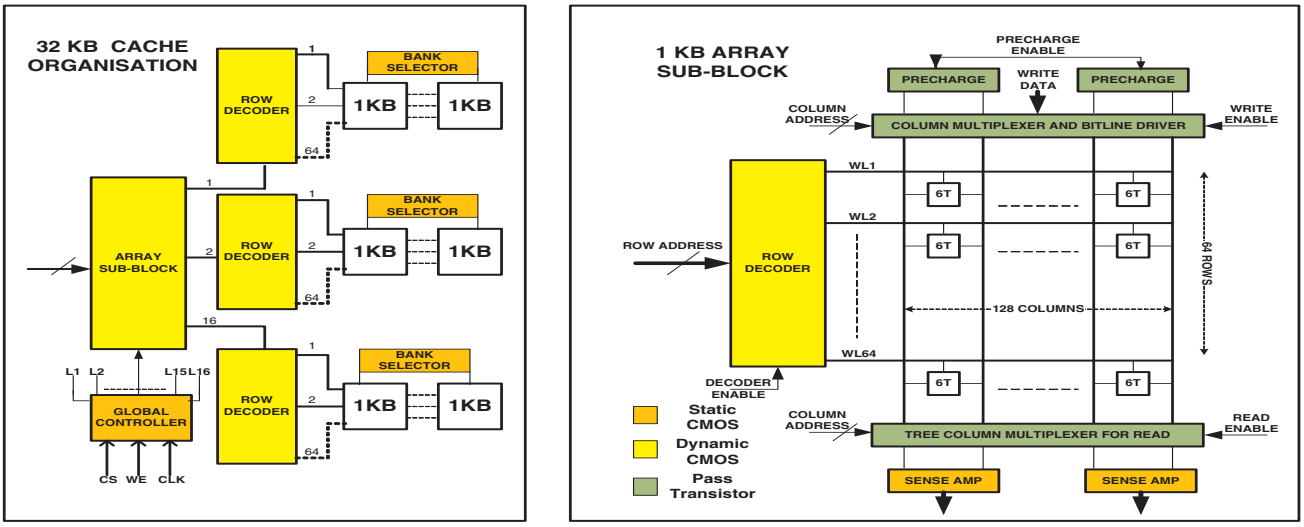

Fig. 5. Cache Architecture

variations with temporal variations. Moreover, we are interested only in the calculation of normalized propagation delay which would give a better idea about worst case performance under ST variations.

Algorithm 1: Compute Delay-Model

Input: Input Netlist and Parameter Distribution
Output: Variation Aware Normalized-Delay Model
while *Netlist Confirms with Floorplan* **do**
  while $NumberofSegments \neq 0$ **do**
    for *Every Monte Carlo Run* **do**
      for *Every Parameter in the List* **do**
        $P = P_{nom} + \delta P_{variation}$;
        Print P;
      end
      Perform Simulation;
      Measure Propagation Delay;
      Normalize Value to smallest Delay
    end
    for *Every Temperature Range* **do**
      Build Lookup Table for Parameter Values and Propagation Delay;
      Fit the Input Parameters to Normalized Output Delay using Model;
      Extract the Slopes of the Model;
    end
    Build Lookup Tables for Slope Values and Temperature Values;
    Fit Input Temperature to Output Slopes using Linear Regression;
    Print Model Parameters;
  end
  $NumberofSegments = NumberofSegments - 1$
end

Extending the model proposed in the Section 4 to calculate the access time would be a cumbersome process due to the existence of very long paths from the control to the cell(sense amplifier) for write(read) access measurements. Looking at the model for R/W access in Equation 10&11, we find that the access time is the summation of individual segment delays where each segment delay is a dependent variable. This analysis is a specific form of the Multivariate Analysis of Variance (MANOVA) [22] where the interaction between the independent variables of every segment (L_{eff},V_{th} and temperature) with its dependent variable (segment delay) is established and the individual contribution to the final dependent variable (access time) is determined. There are two main advantages of breaking the logic path into smaller segments. Firstly, it would greatly reduce the number of paths and secondly, the estimation of segment delay would reduce the error calculation in the overall delay.

VI. Model Validation

A. Experimental Setup

The cache is simulated on the HSPICE simulator using 16nm Predictive Technology Model [23], [18]. The parameter deviations

TABLE I
PARAMETER DEVIATION

Parameter	D2D σ/μ	WID σ/μ_{sys}	WID σ/μ_{rand}
V_{th}(nmos,pmos)	±3%	±6.4%	±6.4%
L_{eff}	±3%	±3.2%	±3.2%
Temperature	30°C-110°C(Step 10°C)		

at each level of granularity is specified in Table I. In order to provide scope for a broad design space, the estimates made in TableI are large enough to enclose any reasonable design point. From [16], we know that the standard deviations of both systematic and random components are equal and σ/μ of L_{eff} is strongly correlated to the systematic component of V_{th}. For every temperature interval we run 500 Monte-Carlo samples. This is sufficient enough to validate the assumption that propagation delay is linearly dependent on the selected input parameters. The monte-carlo samples generated for each temperature interval is constant across the entire range. The access times are normalized to the fastest cache at 30°C so as to facilitate the estimation of worst, best and average case behavior. The simulations were performed on a machine with a dual-core processor running at near 3GHz with 4GB of main memory.

B. Simulation Results

Figures 6-11 show the error box-plots between our Model output and the HSPICE simulation output. The error is computed for every individual segment present in the access time model. The advantage

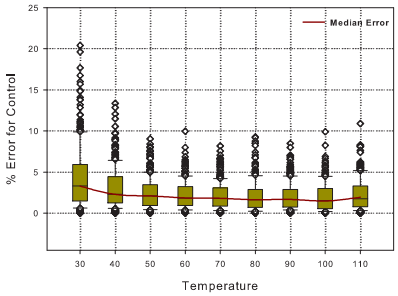

Fig. 6. Control Circuitry

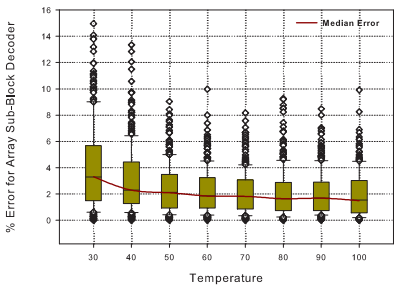

Fig. 7. Array Sub-Block Decoder

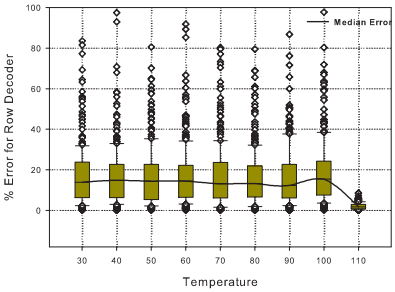

Fig. 8. Row Decoder

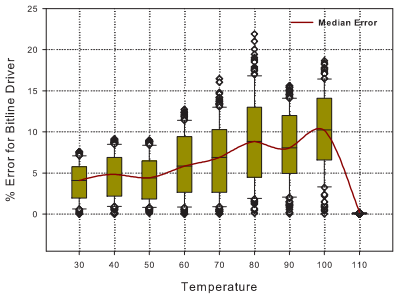

Fig. 9. Bit Line Driver

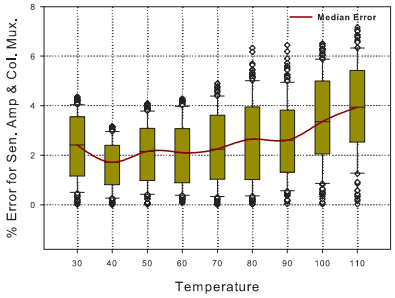

Fig. 10. Sense Amp. & Column Multiplexer

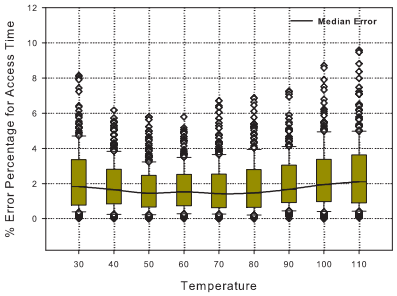

Fig. 11. Access Time

Error in Calculation of Propagation Delay between Proposed Scheme and HSPICE

of such a segment based methodology was to test the model over different types of structures (static, dynamic and pass). In five cases the median error computed is less than 5% with the lowest being 1.8%. In the specific case of the address decoder, the model performance is degraded as the median error is of the order of 14%. This can be attributed to the fact that the address decoder built with dynamic CMOS gates drives a large load inside the sub-block. Thus, the effects of ST variation will be most felt by this structure and the non-linear behavior is very tough to be captured by such a linear model. However, due to a segment based model computation, the overall error in the model can be reduced substantially with higher number of Monte Carlo runs.

There is no relation between the increasing temperature and the error rate. At higher temperatures due to the very high dependence of delay on temperature, access time failures [19] occur very frequently. It results due to the increase in read/write times. Thus the number of successful samples generated for post-simulation analysis will decrease with every temperature increment. This phenomena can be particularly observed in the case of sense amplifier & column multiplexers which drive large loads with minimal number of devices.

Table II presents the simulation time for HSPICE and the model. As there is a linear dependence with the number of runs, the model achieves very high speedups with increasing number of samples. For instance, when running a decent amount of samples (over 500 simulations), the speed-up of the model is of the order of $1x10^5$.

TABLE II
SIMULATION TIME SPEEDUP

Monte Carlo Runs	HSPICE Execution Time(s)	Model Execution Times(s)	Speed Up [$x10^3$]
50	7105	0.245	29
100	14960	0.312	47.9
200	30304	0.378	80.2
500	72243	0.557	129.7
1000	143757	0.978	146.9
2000	289260	1.762	164.2

VII. CASE STUDY

A. Case Study I

This section uses the model derived and validated in the previous sections to study the effect on the access time of a 32KB direct-mapped cache memory under process and temperature variations. To conduct the study, we took a sample of 500 dies. Figure 12, shows the

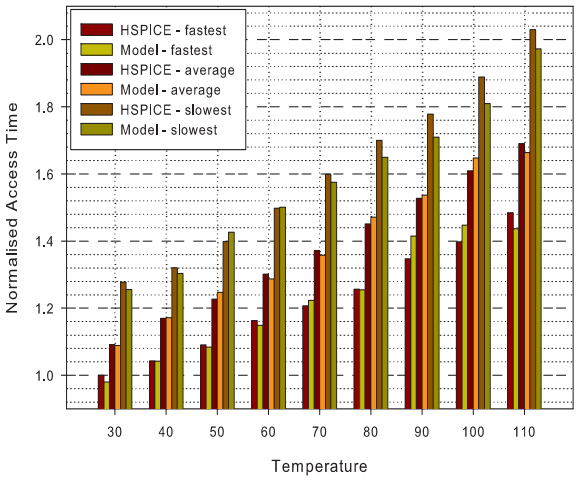

Fig. 12. Variation of Normalized Access Time across Different Dies for Different Temperatures.

normalized access time for the fastest, median and slowest chip for each temperature. At the same time, we also plot HSPICE simulations on those chips so that the precision of the model can be seen for each configuration and across temperatures. In the case of best and worst case performance, the model results are the equivalents of the simulator output. For average performance, the average over the entire set of available output samples was calculated for both simulator

and model outputs. In accordance with the linear dependence of delay with temperature, the access time increases sharply beyond temperatures of 60°C. Comparing to Figtre 3 (where no process variation is simulated) we can see that when considering process variations, the slowdown on access time that was around 1.6x for 16nm scales up to 2x in the presence of process variations in the worst case. At the same time, given the same temperature, process variations induces a 33% difference in propagation delay between slowest and fastest chips. The combined effect of process and temperature variations degrades significantly the performance of the cache studied which warns us of the negative effects of variations. Plus, using the model proposed in this paper, the study just needed half a second to be completed. This speed and preciseness may allow this model to be a good technique to be incorporated into early stage design tools as well as run-time performance prediction or risk prevention tools.

B. Case Study II

Though the model is capable of considering the impact of parameter variations on circuit propagation delays; it can also be used to evaluate, for instance, the cache behavior under Dual-V_{th} assignments (where only V_{th} varies) [24]. While reducing the V_{th} decreases propagation delay, it increases leakage power. In [19], failures due to variation of threshold voltage within the 6T SRAM cell is discussed. Hence, we run our model (and HSPICE simulations to compare the results) on a cache where the peripheral circuitry has a lower V_{th} than that of the drivers and the memory array. As shown in Figure

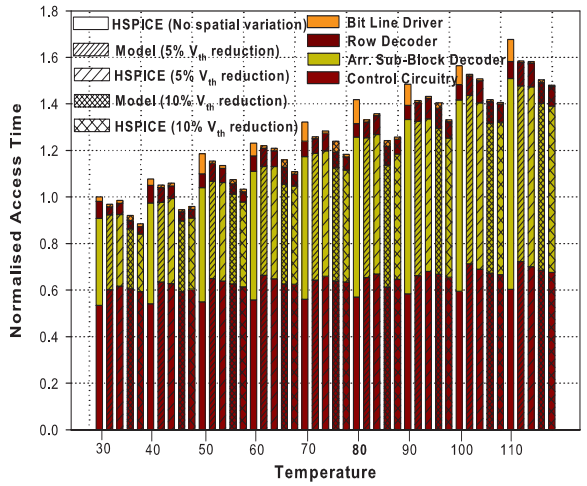

Fig. 13. Impact of Dual V_{th} Assignment on Cache Access Time

13, the write access time is calculated for a 5% and 10% reduction in V_{th} of the peripheral circuitry. This study assumes a single σ/μ for the entire peripheral circuitry. This is similar to assuming a corner case situation when all the components have similar distribution. While at low temperatures the benefits of reducing the threshold are minimal, at higher temperatures for a 10% reduction in V_{th}, the delay is reduced by nearly 18%. As in the previous study, the median error is below 2% and the speedup achieved for generating each of the nine-sets is of the order of $1.3x10^3$.

VIII. Conclusion

In this work, we have proposed an efficient multivariate regression based modeling technique to capture the effects of spatio-temporal variations on propagation delay. In addition, we have extended the current multi-level Quadtree based modeling technique by incorporating layout-geometry based correlation. We then evaluated the model on an in-house built, state of the art 32KB cache to determine the access time deviation due to spatial variation in the presence of temporal variations. Our model is very much scalable with respect to number of input parameters and accurate in terms of calculation of propagation delay. Apart from studying the consequences of spatio-temporal variations on circuits, it can speed-up the forming of heuristics for speed binning and timing yield prediction, enable a fast and accurate design space exploration or even be integrated into a system or architectural simulator used for runtime predictions and risk assessment.

IX. Acknowledgments

This work has been partially supported by the Spanish Ministry of Education and Science under grant TIN2007-61763 and TEC2008-01856, the TRAMS project of the FP7 program of the European Commission under agreement 248789, the Generalitat of Catalunya under grant 2009SGR1250 and Intel Corporation.

References

[1] Horowitz *et al.*, "How scaling will change processor architecture." *ISSCC*, 2004.
[2] Zeitzoff *et al.*, "A perspective from the 2003 ITRS: MOSFET scaling trends, challenges, and potential solutions." *CDM*, 2005.
[3] Bernstein *et al.*, "High-performance cmos variability in the 65-nm regime and beyond." *IBM-JRD*, 2006.
[4] Stine *et al.*, "Analysis and decomposition of spatial variation in integrated circuit processes and devices." *TSM*, 1997.
[5] Chang *et al.*, "Using a statistical metrology framework to identify random and systematic sources of intra-die ILD thickness variation for CMP processes." *IEDM*, 1995.
[6] Fitzgerald *et al.*, "Analysis of polysilicon critical dimension variation for submicron CMOS processes." *MIT, Dept.of ECE*, 1994.
[7] Yu *et al.*, "Manufacturability evaluation of deep submicron exposure tools using statistical metrology." *ISSM*, 1995.
[8] Chen *et al.*, "Modeling and testing of SRAM for new failure mechanisms due to process variations in nanoscale CMOS." *VTS*, 2005.
[9] J. H. Cha *et al.*, "An extended model for optimal burn-in procedures." *TR*, 2006.
[10] Agarwal *et al.*, "Statistical timing analysis for intra-die process variations with spatial correlations." *ICCAD*, 2003.
[11] Xie *et al.*, "False path aware timing yield estimation under variability." *VTS*, 2009.
[12] Miskov *et al.*, "Process variability-aware transient fault modeling and analysis." *ICCAD*, 2008.
[13] Cline *et al.*, "Transistor-specific delay modeling for ssta." *DATE*, 2008.
[14] Visweswariah *et al.*, "First-order incremental block-based statistical timing analysis." *TCAD-ICS*, 2006.
[15] Kumar *et al.*, "Reversed temperature-dependent propagation delay characteristics in nanometer cmos circuits." *TCS*, 2006.
[16] Sarangi *et al.*, "VARIUS: A model of process variation and resulting timing errors for microarchitects." *TSM*, 2008.
[17] Yingmin *et al.*, "Performance, energy, and thermal considerations for smt and cmp architectures." *HPCA*, 2005.
[18] "Predictive Technology Models, http://www.eas.asu.edu/ ptm."
[19] Agarwal *et al.*, "A process-tolerant cache architecture for improved yield in nanoscale technologies." *TVLSI*, 2005.
[20] Liang *et al.*, "Mitigating the impact of process variations on processor register files and execution units." *MICRO*, 2006.
[21] Wilton *et al.*, "CACTI: An enhanced cache access and cycle time model." *JSSC*, 1996.
[22] E. Çinlar, "Introduction to Stochastic Processes." 1975.
[23] W. Zhao and Y. Cao, "Predictive technology model for nano-cmos design exploration." *JECTS*, 2007.
[24] Liu *et al.*, "Leakage power reduction by dual-vth designs under probabilistic analysis of vth variation." *ISPLED*, 2004.