A Novel Fault Localization Technique Based on Deconvolution and Calibration of Power Pad Transients Signals

Reza M. Rad* and Jim Plusquellic+

Department of CSEE, University of Maryland, Baltimore County
Department of ECE, University of New Mexico

Corresponding Author: Reza M. Rad

Address:

University of Maryland Baltimore County

ITE 374, CSEE Dept.,

1000 Hilltop Circle,

Baltimore, MD 21250.

Phone: 410-455-8776.

E-mail: reza2@umbc.edu.

Fax: 410-455-3969.

A Novel Fault Localization Technique Based on Deconvolution and Calibration of Power Pad Transients Signals

Reza M. Rad* and Jim Plusquellic+
Department of CSEE, University of Maryland, Baltimore County
Department of ECE, University of New Mexico

Abstract

This paper describes a new fault localization method that is based on the analysis of power supply transient signals. Impulse response functions derived from the power grid are used to deconstruct the measured power port transient signals into a set of gate level transients generated by the logic gates as signals propagate along paths in the circuit. By comparing these gate transients with those obtained from a defect-free chip or simulation model, it is possible to identify anomalies produced by defects and to locate them to specific path segments in the layout. Impulse response functions are used to significantly reduce both the attenuation effects of the power grid on the gate-generated transients and the chip-to-chip impedance variations in the power grid and test environment. Non-linear calibration techniques are proposed to reduce the chip-to-chip variations in path delays introduced by process variations. The procedure is demonstrated using simulation experiments to locate the position of defects to one or a small group of gates.

1.0 Introduction

Diagnosis is the process of identifying the location of the fault in chips that have failed in the field or at production test. It is a key component of failure analysis. The information gleamed from failure analysis is used to tune the fabrication process for the purpose of improving reliability and yield.

Hardware-based fault localization is challenged by increases in chip complexity as well as

additional interconnection levels and the limitations on the spatial resolution of imaging technology [1]. The increase in difficulty and cost of performing hardware physical failure analysis is likely to move it into a sampling/verification role. These trends continue to increase the importance of developing alternative software-based fault localization procedures, particularly those that are suited for volume diagnosis.

In this paper, we describe a novel software-based fault localization method that is based on the analysis of a chip's transient signals. The basic objective of the method is to derive the shape of the transient signals generated by the individual gates or groups of gates along sensitized paths from the composite transient signals measured at the power ports or C4s of the chip. Once derived, the individual transients of the gates can be compared with those produced by a defect free chip or simulation model to identify the position of a defect.

There are several challenges to overcome in order for this type of approach to be useful. The first is defining a transformation process that decomposes the measured transient waveforms on the C4 power ports of the chip. Our decomposition process is able to derive the individual gate transients that are produced at the sources of the p-channel transistors as a signal propagates along a path. Moreover, it is able to accomplish this under the condition that multiple gate level switching events are occurring simultaneously. This is important because it is difficult to generate test vectors that propagate signals along only one path in the circuit. The decomposition of multiple, spatially distinct transient sub-events is accomplished by solving a set of n simultaneous equations, one for each of the simultaneously switching gates. The parameters of the n equations are derived from measurements taken from n C4 power ports.

A second challenge involves providing sufficient temporal resolution so that the transient of each gate or small groups of gates can be resolved. This allows the position of the defect to be iso-

lated to a specific region along a path. Since the C4 transient waveforms reflect the entire propagation of signals along paths in the circuit, it is necessary to partition the C4 measured transients into multiple time segments. The set of C4 transients in each distinct time interval are then used as input to the spatial decomposition process to obtain the gate-level transients. The combination of both temporal and spatial decomposition of the C4 transients maximizes the level of diagnostic resolution.

A third challenge is dealing with the effects of process variations. Each chip from which the C4 transients are measured will be different because of process variation effects occurring within the power distribution system, transistors and interconnect. Therefore, in order to maximize the effectiveness of resolving anomalies introduced by defects, it is necessary to calibrate for these adverse effects. In this work, we propose and demonstrate several calibration technique that are able to significantly reduce the signal variations caused by process and environmental variation effects.

The organization of the paper is as follows. Section 2.0 describes other power-signal diagnostic techniques proposed in the literature. The fault localization method is described in Section 3.0. A diagnostic flow is presented in Section 4.0 and the results of simulation experiments are reported in Section 5.0.

2.0 Background

I_{DD} diagnostic techniques can be broadly partitioned into two groups; those based on I_{DDQ} and those based on I_{DDT}. I_{DDQ}-based methods, e.g., [2] through [15], have been effective in the past for detecting and locating shorting defects but are increasingly less effective given the leakage current trends over the last decade. Although I_{DDQ} is still relevant for low power ICs, alternatives are needed for many ASICs and microprocessor/microcontroller-based chips. In previous work,

we proposed the use of multiple supply pad I_{DDQ} measurements as a means of reducing the impact of large background leakage currents and developed a triangulation-based method for localization [16, 17, 18]. This strategy effectively dealt with the leakage problem. However, other deficiencies of I_{DDQ} such as the inability to detect and diagnose delay defects, such as resistive open defects, remain.

An alternative approach to I_{DDQ} is to measure and analyze the chip's transient power supply signals, I_{DDT}. I_{DDT} techniques are robust to increases in leakage current and can be applied to virtually any type of logic, including dynamic logic. Moreover, I_{DDT} techniques have the potential to detect and locate defects that affect delay, such as resistive opens. Several I_{DDT}-based diagnostic techniques have been proposed in the literature.

A fault localization method is proposed in [19] that analyzes the area under I_{DD} waveforms as a means of extracting delay information. This delay information is used to estimate the "logic depth" at which the fault occurs. The authors indicate that the possibility of aliasing may reduce the resolution of the technique in random logic. In [20], a fault localization technique is proposed that is based on the charge delivered to the chip during a transition. The input sequence and corresponding circuit activity can be used to isolate the fault to a region of the chip. This technique can be shown to work well for highly controllable structures such as memories but may be more difficult to apply in random logic. Bhunia et al. [21] analyze the power supply transient signals using a Wavelet transform. Defect localization is achieved by mapping the time at which the wavelet transform coefficient of the defective chip differs from that of the defect-free chip into logic depth.

The main deficiency of these techniques is related to resolution. The application of a stimulus, i.e., a two-vector sequence, produces a large transient pulse as the clock fires and multiple signals

propagate along different paths in the chip. Therefore, a large fraction of the transient signal generated on the power supply is background noise. I_{DDT} methods must filter this background noise in order to be effective. A second deficiency is related to process and environmental variations. Process and environmental variations introduce signal variations in chips that are otherwise identical. These signal variations can be significant and must be calibrated for.

The techniques described in this paper deal with these deficiencies directly. By measuring multiple power port signals, the adverse effects of a large background noise component are reduced. Signal calibration techniques, such as those described in [16, 17, 18], are proposed for dealing with process and environmental resistance variation effects. These techniques are extended in this paper to calibrate for impedance variations.

3.0 I_{DDT} Diagnostic Process

Our techniques are based on the analysis of a chip's transient signals. The basic objective of the method is to derive the shape of the transient signals generated by the individual gates along sensitized paths using the composite transient signals measured at the power ports or C4s of the chip. Once derived, the individual transients of the gates can be compared with those produced by a defect-free chip or simulation to identify the location of a defect. The main steps of the proposed I_{DDT} diagnostic process are: 1) path extraction/simulation, 2) power grid characterization, 3) signal calibration, 4) temporal decomposition, 5) spatial decomposition. Each of these is described in the following sections.

3.1 Path Extraction/Simulation

The first step is to identify the candidate paths from the production test fail data. The paths are obtained by simulating the failing test pattern sequences using a logic-level simulator. Once the

candidate paths are identified, the defect-free transients produced by the gates along these paths are obtained from a SPICE simulation of a RC-transistor model. In order to keep the simulation tractable, the RC-transistor model includes only those paths that are sensitized, i.e., the power grid and gates belonging to other non-sensitized paths are not included in the simulation model.

A method for extracting and simulating paths is proposed in [22]. An important element of the process is to preserve the coupling capacitors from these paths to the power grid and to paths that are not included in the model. Otherwise, the transients produced from the simulation of the sensitized paths do not reflect the transients that would have been produced under the full chip simulation model (and do not accurately represent the hardware). This requires that the path extraction process be performed on the full chip RC-transistor model. The coupling capacitors along the paths of interest to other elements in the full chip model can be identified and extracted (and tied to GND) in the smaller RC-transistor model used in the simulations. This model for extraction represents a departure from the procedure proposed in [22].

We do not perform the path extraction/simulation portion of the diagnostic flow in this work because the focus here is on proof-of-concept. Future work will investigate the application of the method to a commercial CUT, in which the path extraction procedure will play an important role.

3.2 Power Grid Characterization[1]

The power grid can be modeled as a linear, time-invariant (LTI) system that acts as a low pass filter for transient waveforms[2]. The ***inputs*** to the LTI system are the gate-level transients produced on metal 1 as signals propagate along sensitized paths and the ***outputs*** are the transients produced

---

1. Power grid characterization is previously discussed in [22] and briefly discussed here for ease of reference.
2. Current transients are analyzed in this paper, but the same model and procedure holds for voltage transient analysis.

on the C4 power ports of the chip. Under these assumptions, it is possible to characterize the LTI system using a set of impulse response (IR) functions, one between each input (gate source connection) and output (C4 power port). Once the IR functions are known, they can be used to predict the C4 transient waveforms produced under any arbitrary input stimulus using convolution and superposition. Likewise, employing deconvolution and solving simultaneous equations allows the reverse process, i.e. given C4 transients as input, it is possible to predict the gate level transients. This latter process is the objective of our diagnostic procedure.

The LTI system describing the power grid is a multi-input, multi-output system. The set of IR functions, which define the relationships between the inputs and outputs, can be obtained from simulation experiments as shown in Figure 1. Here, an RC model of the power grid is constructed (displayed as a mesh in the figure)[1], with inputs defined in metal 1 as the positions at which the gates attach to the power grid in the layout (only one position, labeled *node 1*, is shown in the figure).

The outputs are the C4 power ports, labeled $C4_1$ through $C4_4$ along the top of the figure. In each simulation experiment, a current source is attached to one of the inputs and a current step is driven onto the grid. The step responses measured at the outputs are converted to IR functions by differentiation. For the configuration in Figure 1, four IR functions are computed for the input defined as *node 1* labeled h_{11} through h_{41}. These functions are a key component of deconvolution as described in the following sections.

The power grid characterization process differs depending on whether the diagnostic procedure is to be applied to a small number of chips or whether it is to be used for volume diagnosis,

1. Only one RC model of the power grid needs to be extracted from the layout, e.g., using the nominal values of the process. Process variations are handled by a separate procedure.

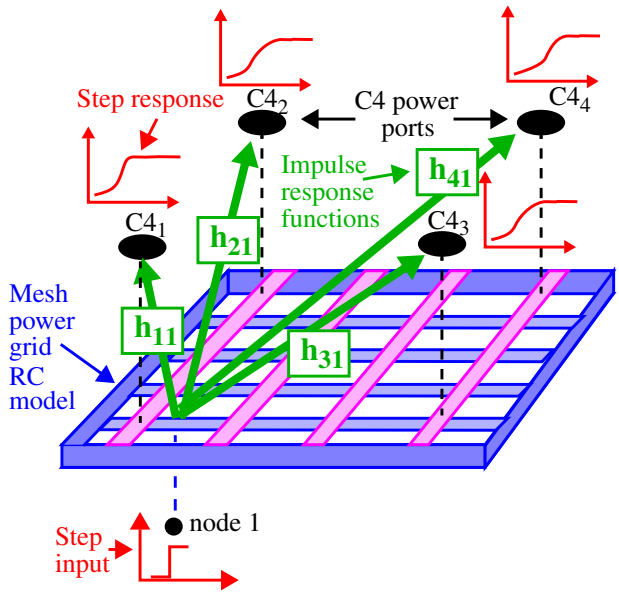

Fig. 1. Power grid characterization from substrate connections to C4s using step signals.

where a large number of defective chips are processed.

3.2.1 Small Sample Diagnosis

In situations where only a few CUTs need to be diagnosed, the power grid characterization described above is performed only for a subset of the gate connections in the simulation model of the CUT. This avoids a time consuming complete characterization of the power grid, where, for example, one simulation is carried out for each gate in the core logic. In this case, the connection points for which the IR functions are computed are those that correspond to gates belonging to a set of candidate paths, as discussed in Section 3.1.

3.2.2 Volume Diagnostics

For volume diagnostics, completely characterizing the power grid can be more efficient than the process described above. In this case, the current step input simulation is repeated for all gate connections in the core logic. For a power grid with n inputs and m outputs, n transient simulations are required and nxm IR functions are computed. If the RC model of the power grid and/or

the number of gates (inputs) in the chip is large, this could result in significant computational cost. Two characteristics of the LTI system defining the power grid can be leveraged to reduce this cost significantly.

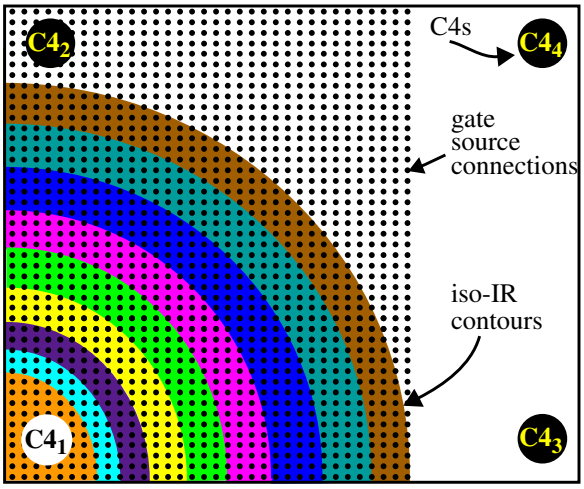

Fig. 2. Iso-IR contours of a power grid.

First, as demonstrated in previous work, the IR functions for inputs in concentric bands, called iso-IR contours, around the C4s are nearly the same, and therefore can be shared [22]. This characteristic is illustrated in Figure 2, which gives a top-down view of a power grid with four C4s. The IR functions are computed between each of the gate connections (inputs), shown as small dots, and $C4_1$ (output) given in the lower left-hand corner. A similarity metric described in [22] is computed for each IR function with respect to its neighbors. The inputs with IR functions that are within, e.g., 5% according to the metric, are shown within a set of iso-contour bands in the figure (only the inner most contours are shown). Therefore, a single IR function can be shared among all inputs within each contour band. This feature of a power grid significantly reduces the number of inputs that need to be simulated because any one of the inputs within the band can be used to represent the other inputs in the band.

Second, the entire RC model of the power grid is not needed in each simulation. In previous work, we showed that the impedance to C4s that are not in the immediate vicinity of the input

increases quadratically as the distance between the input and C4 is increased [23]. Therefore, the IR functions to C4s outside the immediate region of the input are of no consequence because of the large impedance between the input and output. By eliminating these C4s and corresponding RC components of the power grid from the simulation model, the size of the RC model is significantly reduced, and more easily simulated.

3.3 Signal Calibration

The third major step associated with our I_{DDT} diagnostic procedure involves calibrating the measured transient signals to reduce or eliminate the effects of process and environmental (PE) variations. PE variations 'distort' the CUT's transients, making it more difficult to determine the gate that is defective.

The objective of calibration is to make it possible to compare the C4 transient signals obtained from simulations (the reference) with those measured from the CUTs. It is possible to calibrate the reference to the CUT or vise versa. In this work, there is no advantage either way so we choose the former.

The calibration methods that we propose calibrate for two sources of PE variations; linear variations in the power grid and test environment and non-linear variations in the propagation of signals along logic paths. For linear variations, a signal transformation procedure, which uses IR functions similar to those described in the previous section, can be used to deal with these efficiently. Here, the idea is to run a set of calibration tests on both the simulation model and the CUT. The calibration tests performed on the CUT require the insertion of special calibration circuits that are exercised during these tests. The data collected from the calibration tests is used to define one or more transfer functions that are used to calibrate subsequently measured diagnostic test data from the CUT. This procedure is described in the next sub-section.

The linear calibration procedure 'fixes' signal distortions introduced by variations in the RLC components of the power grid and test environment. However, it does not handle differences that occur in signal propagation among the CUTs and reference. Process variations in transistor components, including channel length, channel doping and threshold voltage change the transistor switching characteristics in a non-linear fashion. These switching variations introduce variations in the individual gate transients as signals propagate along sensitized paths under an applied test sequence. This, in turn, affects the transient signals measured at the C4s of the CUT. Similar to linear variations, non-linear C4 transient signal variations need to be calibrated for (removed) to enable a meaningful comparison to be made between the CUT and reference.

3.3.1 Calibrating Linear Variations

The calibration of linear PE variations can be accomplished using the calibration circuit (CC) as shown in Figure 3(a). A similar calibration structure is proposed in [16] for I_{DDT} signals and in [17] for I_{DDQ} signals. The calibration circuit is composed of a scan flip-flop, whose output is connected to an n-channel transistor. A current step, similar to that shown in Figure 1, is introduced by scanning a '1' into the flip-flop. The IR functions for these layout positions are computed as described in Section 3.2. This process is carried out on each of the CUTs and on the reference simulation model of the power grid.

We investigate three calibration schemes in this work. In the first two schemes, the CCs are distributed across the layout of the chip, as shown in the center of Figure 3(b)[1]. Calibration schemes 1 and 2 differ only in the number and granularity of calibration circuits used. For the third approach, a matrix is constructed based on the four CCs that are placed underneath the C4s, also

1. The scan chain is not shown in Figure 3(b). It connects the calibration circuits used in each approach together to implement a shift register.

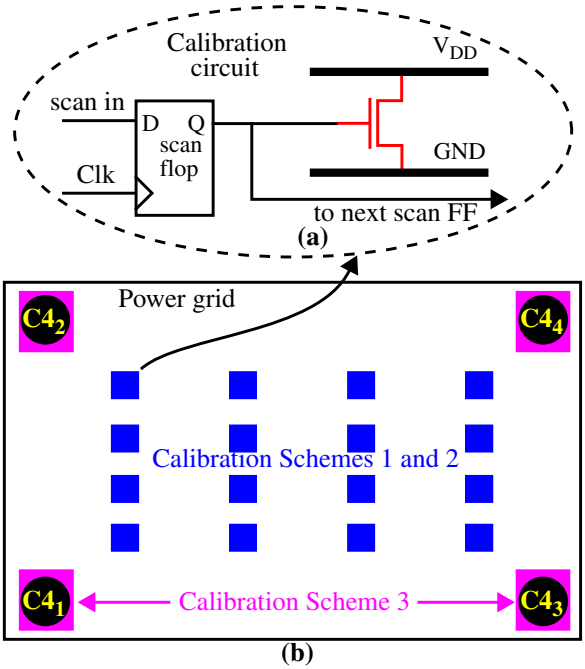

Fig. 3. (a) Calibration circuit and (b) the three calibration schemes.

shown in Figure 3(b).

- ***Distributed Calibration Circuit Approach***

The general approach to calibration for schemes 1 and 2 is given as follows. Prior to calibration, the layout positions of the gates along failing paths for a CUT are determined as well as a set of CCs that are topologically closest to these gates. These CCs are used to derive the transfer functions needed for calibration.

The calibration process for one of the gates of a candidate path is shown in Figure 4 with the simulation-derived reference on the left and the CUT on the right[1]. An arrow along the bottom of the power grids indicates the connection point of the gate. We refer to this layout position using the coordinates (x_1, y_1). The closest calibration circuit (CC), identified from the array shown in Figure 3(b), is also displayed along the bottom of the power grids. We refer to this layout position using the coordinates (x_2, y_2).

1. This process is repeated for other gates along the path.

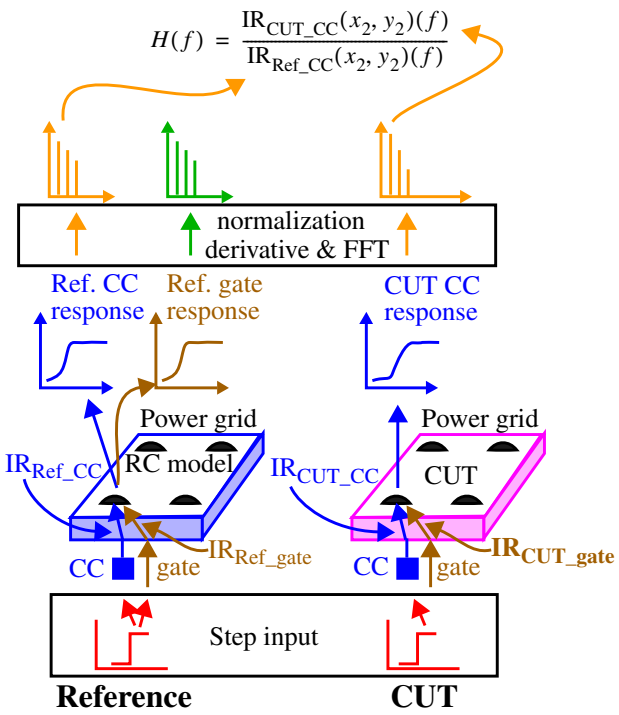

Fig. 4. Calibration procedure for distributed calibration circuit approach.

The transfer function is constructed by performing a sequence of tests in the reference and CUT. Current steps are first applied using the CCs at layout positions (x_2, y_2) and the step responses measured from the C4s. The step responses for $C4_1$ are shown in the figure for both the reference and CUT. The step responses are then *normalized* by dividing each point by the total charge drawn by the current step over the time interval of the measurement. The total charge is easily computed as the sum of the areas under the step response curves at all C4s[1]. A similar normalization approach is performed in [16].

The derivative of the normalized step response yields the IR functions. The time domain IRs of the reference and CUT are converted to their frequency domain representations using a FFT and *H(f)* is defined as their ratio (shown along the top of the figure). *H(f)* defines a transfer function

1. The normalization process accounts for current step variations that occur in the CCs of the CUT due to process variations.

that effectively nulls out the impact of PE variations in the CUT.

H(f) cannot be used directly to eliminate PE variations because it is computed from the CC at position (x_2, y_2). The gate is actually connected to the power grid at a nearby point (x_1, y_1). Therefore, in order to obtain a better estimate of the (x_1, y_1) transfer function, a third test is performed in the reference. A step input is configured in the reference to drive the power grid at the gate's connection point (x_1, y_1), as shown on the left side of Figure 4. The step responses at the C4s are processed in the same manner to derive the IR responses, identified as IR_{Ref_gate} in the figure. This IR function in combination with *H(f)* computed above can be used to estimate the CUT's IR function, IR_{CUT_gate}, using Equation 1.

$$IR_{CUT\_gate}(x_1, y_1)(f) = IR_{Ref\_gate}(x_1, y_1)(f) \times H(f) \quad \textbf{(1)}$$

The IR functions for other gates along the path are computed in a similar fashion. These IR functions are used to derive each of the gate transients produced in the CUT under the diagnostic test using the procedures described in Sections 3.4 and 3.5.

- ***C4 Matrix-Based Approach***

The third calibration scheme that we investigate makes use of a smaller set of CCs, one under each C4 as shown in Figure 3(b). This third scheme is similar to a calibration method proposed in [16] for calibrating I_{DDT} areas and is based on a standard linear algebra transformation process. The general form of the calibration method is extended in this work to handle transient waveforms.

The main shortcoming of the distributed CC calibration method described above is overhead, i.e., a large number of CCs need to be distributed across the chip to obtain a good estimate of a gate's IR function. The C4 matrix-based approach requires only one CC per C4, and therefore, has a much lower overhead. More importantly, the use of information from multiple CCs in the

matrix operation enables an accurate estimation of all IR functions in the 2-D layout space enclosed by the CCs. Therefore, the matrix approach can potentially deal with PE variations more efficiently. Sections 5.3 and 5.4 compare the effectiveness of the distributed CC method with the C4 matrix approach using simulation experiments.

The calibration process is similar to that illustrated in Figure 4 except that the transformation function *H(f)* is a matrix constructed using the data from all CC tests. The IR functions in the matrix are labeled as $IR_{ij}(f)$, with *i* identifying the CC and *j* identifying the C4. For a power grid with four C4s, the reference calibration matrix is defined as a 4x4 matrix, $MIR_{REF\_CC}(f)$, as given by Equation 2.

$$\mathrm{MIR}_{\mathrm{Ref\_CC}}(f) = \begin{bmatrix} IR_{11} & IR_{12} & \dots & IR_{1M} \\ IR_{21} & IR_{22} & \dots & IR_{2M} \\ \dots & \dots & \dots & \dots \\ IR_{M1} & IR_{M2} & \dots & IR_{MM} \end{bmatrix} \quad \textbf{(2)}$$

The calibration tests are carried out on the CUT as well and the corresponding matrix is given by $\mathrm{MIR}_{\mathrm{CUT\_CC}}(\mathrm{f})$ in Equation 3. The transfer function is defined as a matrix product given by Equation 4. The RHS is defined as the inverse of $\mathrm{MIR}_{\mathrm{Ref\_CC}}(\mathrm{f})$ multiplied by $\mathrm{MIR}_{\mathrm{CUT\_CC}}(\mathrm{f})$. The

$$\mathrm{MIR}_{\mathrm{CUT\_CC}}(f) = \begin{bmatrix} IR_{11} & IR_{12} & \dots & IR_{1M} \\ IR_{21} & IR_{22} & \dots & IR_{2M} \\ \dots & \dots & \dots & \dots \\ IR_{M1} & IR_{M2} & \dots & IR_{MM} \end{bmatrix} \quad \textbf{(3)}$$

$$\mathrm{H}_{M \times M}(f) = \mathrm{MIR}^{-1}_{\mathrm{Ref\_CC}}(f) \times \mathrm{MIR}_{\mathrm{CUT\_CC}}(f) \quad \textbf{(4)}$$

IR functions to each of the C4s for an arbitrary gate in the layout at position (x, y) are computed using Equation 5.

$$\mathrm{IR}_{\mathrm{CUT\_gate}}(x,y)(f)_{1\times M} = \mathrm{IR}_{Ref\_\mathrm{gate}}(x,y)(f)_{1\times M} \times H_{M\times M}(f) \quad \textbf{(5)}$$

3.3.2 Calibrating Non-Linear Process Variations

A different approach is needed to calibrate for non-linear variations that occur in path delays among the CUTs and reference model. There are two approaches investigated in this work; a defect-free neighbor path approach and a scaling approach.

- ***Neighbor Path Approach***

The strategy used in this approach is to identify one or more paths in the layout that are close to the path suspected of containing the defect. Tests are derived to exercise these defect-free neighbor paths and a transfer function is constructed using the CUT and reference C4 transients. Although similar to the distributed CC scheme described in Section 3.3.1, this approach can potentially calibrate both linear and non-linear variations.

Figure 5 illustrates the procedure where a set of defect-free neighbor paths, labeled NP_1 and NP_2, are shown using thick lines on either side of the defective path in the CUT. Tests are derived

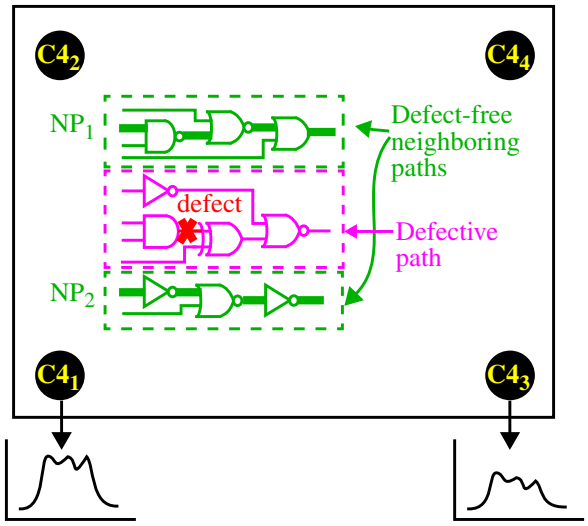

Fig. 5. Calibration procedure for defect-free neighbor path approach.

to propagate transitions along these paths in the CUT and reference. The transfer function $H_{ij}(f)$

given by Equation 6 is defined as the ratio of the frequency domain representations of the CUT and reference transients for C4 *i* and neighboring path *j*. For example, $H_{11}(f)$ is defined as the ratio

$$H_{ij}(f) = \frac{\mathrm{NP}_{Ref\_ij}(f)}{\mathrm{NP}_{CUT\_ij}(f)} \quad \textbf{(6)}$$

of $C4_1$ transients, one measured on the CUT and one obtained from the reference, under a test that sensitizes path NP_1. Given this transfer function, a calibrated estimate of the defective path (DP) transient at $C4_i$, e.g., $DP_{Ref\_i}(f)$, can be obtained using Equation 7.

$$\mathrm{DP}_{Ref\_i}(f) = \mathrm{DP}_{CUT\_i}(f) \times H_{ij}(f) \quad \textbf{(7)}$$

$DP_{Ref\_i}(f)$ represents an estimate of the transient that the path-under-investigation would generate if the CUT has similar process parameters as the reference. If the path-under-investigation does include a defect, $DP_{Ref\_i}(f)$ will be different from that of the reference. Note that $DP_{Ref\_i}(f)$ represents only an estimate because the layout positions of the neighbor and defective paths are different, and therefore the transfer function $H_{ij}(f)$ is an approximation. One approach to improving the estimate is to average the transfer functions computed for more than one neighboring path, e.g., NP_1 and NP_2 in Figure 5, as given by Equation 8. Once $DP_{Ref\_i}(f)$ is obtained, the IR functions of the reference grid can be used in the signal decomposition techniques discussed in Sections 3.4 and 3.5 to obtain the individual gate transients.

$$\mathrm{DP}_{\mathrm{CUT\_i}}(f) = \mathrm{DP}_{\mathrm{Ref\_i}}(f) \times \mathrm{ave}(H_{ij}(f)) \quad \textbf{(8)}$$

- ***Scaling Approach***

The scaling approach is based on the assumption that process variations are uniform across portions of the CUT, i.e., the delays of gates are correlated in 'regions' of the CUT. Regions are defined, for example, as the area *surrounding* and *including* the layout area enclosed by four C4s, as shown in Figure 3(b). Due to the finite impedance characteristics of the power grid, switching

activity that occurs outside this region has only a small impact on the transients within the region of interest. Therefore, we do not need to assume that process variations are global and uniform across the entire CUT, which is not realistic in advanced technologies.

The basic idea is to determine two scaling factors that can be applied to the transients measured from the CUT to account for differences in path delay between the reference and CUT. For example, Figure 6(a) gives the I_{DDT} waveforms generated by a 10-inverter chain under two 0.25 um process models obtained from MOSIS [25], labeled P_1 and P_2. The same modeling parameters were used for all inverters in each of the chains (uniform regional variation). It is clear that the delay of the chain under P_2 process parameters is slower than the delay under P_1. However, the shape of the waveforms is preserved. This is confirmed in Figure 6(b) where we have applied two scaling factors to the P_2 waveform, one for time (x) and one for magnitude (y). Although small difference still exist between the two waveforms, the main deviations in the original waveforms have been eliminated using the scaling factors.

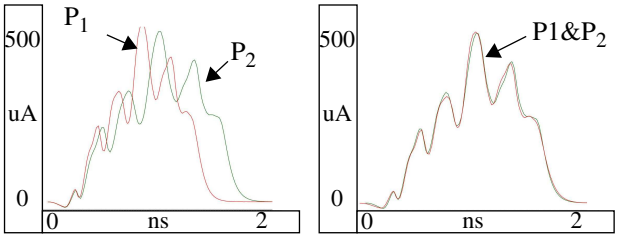

Fig. 6. (a) Original I_{DDT} and (b) scaled I_{DDT} from two process runs.

The scaling factors can be determined in a straightforward manner by scaling the measured C4 transients from the CUT until the difference between them and the C4 transients of the reference are minimized. The metric to minimize is the area under the difference waveform. Binary search can be used to make this computationally efficient.

Under an actual diagnostic test, the C4 transients are the superposition of the gate transients

from more than one sensitized path. In order for this technique to produce good results, it must hold that intra-die process variations are small and therefore, the main effect of process variations is to affect all paths in a similar manner. Although we feel this is a valid assumption for sufficiently small regions in the CUT, as technology is scaled and intra-die process variations get larger, a more sophisticated approach may be necessary.

3.3.3 Noise and Numerical Errors

Calibration is effective for reducing the adverse impact of fixed sources of variations but does not handle noise sources in the test environment or numerical errors in the simulation experiments. In an ideal case, the gate transients can be obtained by dividing the measured power port transients by the impulse responses in frequency domain, as given by the equations in previous sections. In practice, the division operation is sensitive to noise and numerical simulation error, particularly the high frequency components of the transient signals. In this work, we use white Gaussian noise (10 dB signal-to-noise-ratio) to model worst-case noise and numerical error and apply Wiener deconvolution to deal with them [24].

The application of Wiener deconvolution is given as follows. For a LTI system with impulse response IR(t), input x(t) and additive noise n(t), the output of the system, y(t), is given in Equation 9. In order to estimate x(t) using the measured value of y(t) while minimizing the role of the

$$y(t) = IR(t) * x(t) + n(t) \quad \textbf{(9)}$$

noise component, one has to use Equation 10 where g(t) is the Wiener deconvolution filter. The

$$\hat{x}(t) = g(t) * y(t) \quad \textbf{(10)}$$

frequency domain representation of g can be obtained from Equations 9 and 10 and is given in Equation 11. In this equation, N and S are the power spectral densities of noise and signal, respec-

$$G(f) = \frac{IR^{*}(f)S(f)}{|IR(f)|^{2}S(f) + N(f)} \quad \textbf{(11)}$$

tively. Equation 12 re-expresses Equation 11 in matrix format to suit our multiple input, multiple output LTI system. The waveforms produced by solving Equation 12 represent the transients produced by sensitized gates under the test sequence.

$$[X]_{n\times 1} = [G]_{n\times m} \times [Y]_{mx1} \quad \textbf{(12)}$$

3.4 Temporal Decomposition

Temporal decomposition (TD) is the process of breaking up a C4 transient into a set of constituent gate transients. TD is performed in our diagnostic flow on each of the calibrated C4 transients from the CUT. Figure 7 illustrates TD under the condition that the test applied to the CUT sensitizes only one path. The transient measured at the C4, designed as *y*, is shown in the center of the figure. The goal is to derive each of the gate transients, beginning with x_1, shown along the bottom of the figure. The time interval $[t_1,t_2]$ encloses the entire x_1 gate transient. However, gate transient x_2 overlaps this interval and must be accounted for. The portion of x_2 that overlaps this time interval is shaded in Figure 7.

The reference gate transients are used to eliminate the contribution of x_2 from the segment of the C4 transient $y[t_1,t_2]$. The reference gate transient, x_2, is first calibrated for gate delay variations between the reference and CUT (see Section 3.3.2) and is then convolved using the IR function $IR_{CUT}(2)$ to account for linear PE variations. The convolved gate transient is subtracted from the measured C4 transient to eliminate its contribution. The x_1 gate transient is then obtained by deconvolving $y[t_1,t_2]$ using $IR_{CUT}(1)$.

Equation 13 gives the expression used in this process. The C4 transient, *y*, over the time inter-

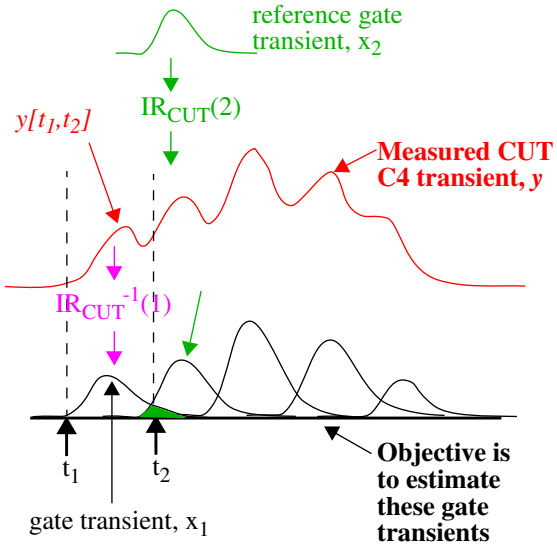

Fig. 7. Even inverter gate transients from a ten inverter chain

val *(t[i-1],t[i])* is the convolution of the impulse response of gate *i* times x_i plus the parameter *b*. Parameter *b* is defined as the sum of the effects of other gate transients during time period *(t[i-1],t[i])*.

$$y(t[i-1],t[i]) = \mathrm{IR}_{\mathrm{CUT}}(i)^{*}x_i(t[i-1],t[i]) + b \qquad \textbf{(13)}$$

$$b = \sum_{j=1,\, j\neq i}^{n} \mathrm{IR}_{\mathrm{CUT}}(j)^{*}x_j(t[i-1],t[i])$$

Once *b* is calculated, the transient of gate *i* can be calculated using Wiener deconvolution as given by Equation 14 where *Y* and *B* are frequency domain representations of *y* and *b* and G_i is the Wiener filter calculated based on impulse response $IR_{CUT}(i)$.

$$\hat{X}_i = G_i(f) \times (Y - B) \qquad \textbf{(14)}$$

Note that this approach assumes that the CUT's x_2 gate transient is defect-free. If it is not, then this process leaves an anomaly in x_1 that may be interpreted as a defect when x_1 is compared with the reference. In this case, the diagnosis algorithm may not predict the precise location of the defect. The simulation results in Section 5.2 illustrate how often this occurs.

3.5 Spatial Decomposition

The assumption that only a single path is sensitized under any given test sequence is not valid in most, if not all, cases. It is much more likely that a test sequence will sensitize multiple, independent paths simultaneously in different parts of the CUT. In such cases, the transients measured on the C4s are the superposition of simultaneous switching events. In order to determine the gate transients in this scenario, a set of simultaneous equations need to be written and solved, one describing each simultaneous switching event. We propose a spatial decomposition process based on deconvolution for this purpose.

The problem is illustrated in Figure 8 using a set of gate transients generated simultaneously from different paths distributed across a portion of a layout surrounded by four C4s. The gate transients superimpose to produce a C4 transient as shown for $C4_1$ in the figure. A similar process occurs simultaneously on the remaining C4s.

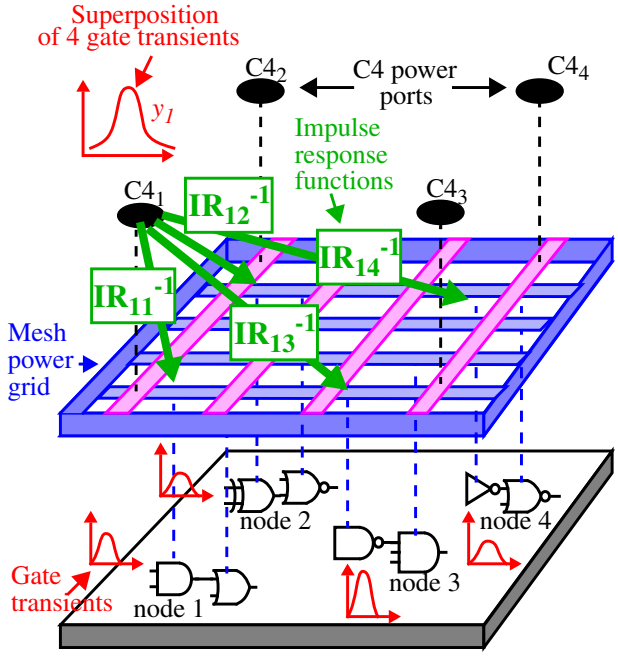

Fig. 8. Exploded chip view with gate transients convolved and superimposed to define C4 transient.

The decomposition of the measured C4 transients is accomplished using the inverse of the IR

functions IR_{11}, IR_{12}, IR_{13},..., IR_{1n}, derived between nodes 1, 2,..., n of the sensitized paths and $C4_1$. We described several techniques in Section 3.3.1 that allows these IR functions to be estimated for the CUT. Given these IR functions, the relationship between the measured transients on the C4s, y_i, and the gate transients along the sensitized paths, x_i, can be described using the convolution relation given by Equation 15, where '*' denotes convolution.

$$y_i = x_1 * IR_{i1} + x_2 * IR_{i2} + x_3 * IR_{i3} + ... + x_n * IR_{in} \qquad i = 1, ..., m \tag{15}$$

Here, n represents the number of nodes or gates that are generating transients simultaneously under the test sequence and m represents the number of C4s from which the transients are measured. Equation 16 gives the matrix form and Equation 17 gives the frequency domain expression for Equation 15. In order to solve for x_i in Equation 15, it is necessary to compute the inverse

$$[y]_{m \times 1} = [IR]_{m \times n} * [x]_{n \times 1} \tag{16}$$

$$[Y]_{m \times 1}(f) = [IR]_{mxn}(f) \times [X]_{nx1}(f) \tag{17}$$

matrix of the IR functions, or G using Wiener deconvolution, as given by Equation 18.

$$[X]_{n \times 1} = [G]^{-1}{}_{n \times m} \times [Y]_{mx1} \tag{18}$$

There are two issues that remain to be addressed. First, it is clear that the systems of equations is not solvable unless $n = m$. The practical consequence of this condition is a restriction on the number of simultaneous switching events, or sensitized paths. In larger chips, there are potentially 100s of C4s for power and therefore, it seems this condition does not pose much of a restriction. However, the impedance characteristics of the power grid impose a restriction on the number of simultaneous events that occur within any given region of the CUT. For example, if sixteen gate transients were shown in Figure 8, this would require the use of transients generated on C4s out-

side this region (not shown). The impedance characteristics of the power grid filter high frequency components of the gate transients, especially on the C4s outside the region. The filtering process adversely affects the resolution of the decomposition process, making it more difficult to resolve the sixteen gate transients. Therefore, for better results, ATPG should be constrained to reduce the number of sensitized paths within any given region of the CUT. We expect the implementation of this constraint to be straightforward because most commercial test pattern generation tools already support constraints on the number of sensitized paths for the purpose of limiting power.

The second issue relates to the temporal analysis given in the Section 3.4. The propagation of signals along any given path in the CUT produces a sequence of overlapping gate transients that must be resolved as well. Figure 9 gives a complete picture of the problem to be solved. One of the measured CUT transients for $C4_1$ is shown along the top of the figure and two path transients are shown below it for paths p_1 and p_2. The path transients are composed of gate transients, which are the target of our decomposition process.

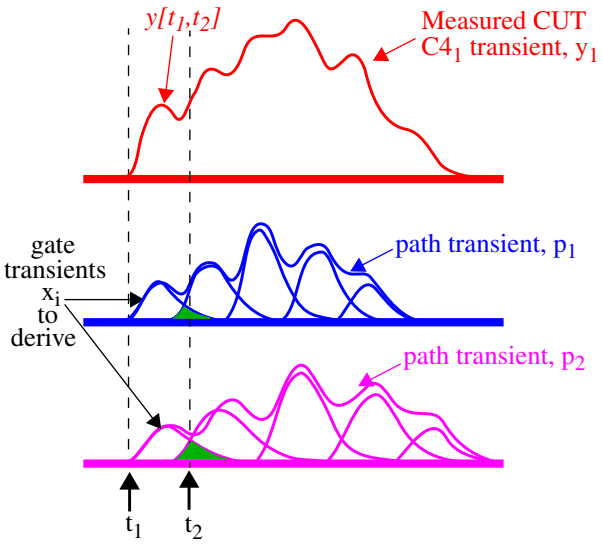

Fig. 9. Temporal and Spatial Decomposition.

For any given time interval $[t_1, t_2]$, the portion of the C4 transient under analysis is given by $y[t_1, t_2]$. Within this interval, the spatial decomposition method can resolve one gate transient on

each path (two in this case), each with a unique IR function as given by Equation 15, using the $C4_1$ transient and one other C4 transient (not shown). However, the output of the spatial decomposition process are the gate transients of interest, identified as x_i in the figure, plus the overlapping transients of the downstream gates driven by the x_i gates. In order to obtain the x_i gate transients, the reference model gate transients are used in a waveform difference operation as described in Section 3.4, except in this case, the operation is performed on each of the sensitized paths.

The general form of the combined spatial and temporal partitioning process is as follows. For any given path, the C4 transients are partitioned in a set of k time segments, each represented by the start and stop points of the gate transients on that path. The spatial deconvolution technique given in Equation 18 is applied using n C4s transients as input (one for each sensitized path). The temporal decomposition process is then performed as given by Equation 19, which describes the relationship between any given C4 transient and the simultaneous switching gate transients over the time interval *(t[i-1],[i])*. This equation is similar to Equation 13 except for parameter *b*.

$$y(t[i-1],t[i]) = \sum_{p=1}^{P} IR(i)^{*}x_i + b$$
$$b = \sum_{p=1}^{P} \sum_{j=1,\, j\neq i}^{n} IR(j)^{*}x_j(t[i-1],t[i]) \qquad \textbf{(19)}$$

Here, b is sum of the effects of all downstream overlapping gate transients from the set of sensitized paths over the time interval *(t[i-1],[i])*. Equation 20 expands on the single path relationship given by Equation 14. This equation is solved for each of the time segments to calculate the gate transients for all sensitized paths.

$$[X]_{n\times 1} = [G]_{n\times n} \times [Y-B]_{nx1} \qquad \textbf{(20)}$$

3.6 Constraints and Limitations

A limiting criteria related to solving Equation 20 is the spatial distribution of the gates under consideration. The IR functions (IR_{ij}s in Equation 15) between C4s and gates must be distinguishable for each gate. In cases in which gates along two paths are too close in the layout, the difference in their corresponding IR functions will be very small and the spatial decomposition method may fail to resolve the individual transients for each gate. This will occur for pairs of converging paths under the test sequence.

In such cases, the gates with very close IRs must be grouped together and considered as a single component in the analysis. Therefore, resolution of the technique will be reduced to small groups of closely located gates rather than a single gate.

4.0 Diagnostic Flow

The proposed I_{DDT} diagnostic can be implemented in a procedure as follows. The extraction and characterization of the power grid of the reference is constant, i.e., independent of the chip under test (CUT), and therefore is carried out before any of the defective CUTs are analyzed. This is given as the first step in Figure 10.

The remaining steps in the process are carried out dynamically, in a manner analogous to a dynamic I_{DDQ} diagnostic method. Under a dynamic paradigm, neither the test set nor the fault database is computed in advance. For our purposes, the test set is derived dynamically from the failing production test(s) that identified the CUT as defective. For example, if the CUT failed one or more stuck-at tests, logic diagnostic techniques can be used to provide a list of candidate sites. For each of these sites, commercial ATPG tools are used to derive a two-vector sequence that propagates a transition to the candidate site. The ATPG tool is configured to choose a vector pair

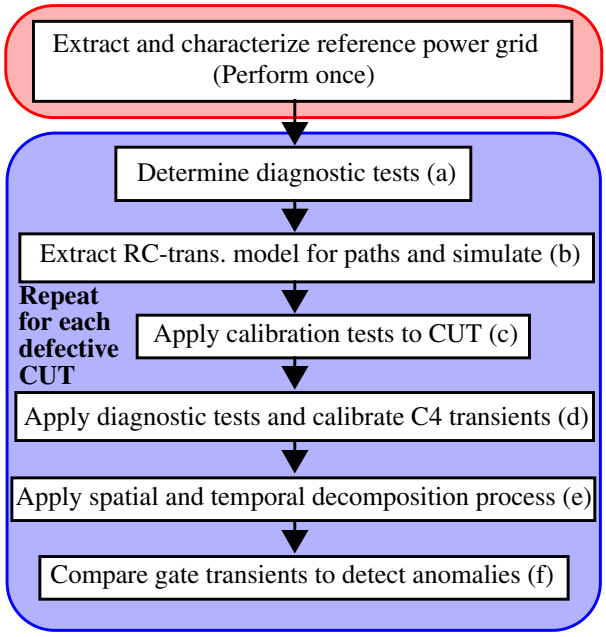

Fig. 10. I_{DDT} Diagnostic Analysis Flow.

such that the number of sensitized paths is as small as possible and the paths are sufficiently separated in the layout. This step is labeled (a) in Figure 10. This addresses the limitations of the method as described in Section 3.6.

For each two-vector sequence, logic simulation is used to determine ALL paths, including path segments, that are sensitized by the two-vector pair (step (b) in Figure 10). The sensitized paths are extracted from an RC-transistor model of the chip and simulated by themselves, as a means of avoiding a full chip SPICE-level simulation. The gate-level I_{DDT} transients are saved as piecewise linear waveforms and represent the 'defect-free' behavior of the chip under this test sequence.

The calibration tests are carried out on the CUT (step (c) in Figure 10). The simulation calibration data and the CUT calibration data are used together to define a transformation process for calibrating for linear and non-linear process variations, as described in Section 3.3. The calibration procedure(s) are used to calibrate the reference C4 transient signals to the CUT as each of the two

vector tests are applied (step (d) in Figure 10). The gate transients of the CUT are derived from the measured C4 transients using the spatial and temporal decomposition processes described in Sections 3.4 and 3.5 (step (e) in Figure 10).

In cases in which the test sequence provokes the defect, an anomaly in a gate transient will be present. Anomalies are identified by comparing the gate transients of the reference to those derived for the CUT (step (f) in Figure 10). There are a variety of metrics that can be used to determine if the gate transient waveforms from the CUT and reference match. The metric used here accounts for measurement noise and other error sources by defining a threshold. The gate transient waveform parameters compared are the areas under the waveforms, the peak value of the waveforms and time position of the peak values. A threshold of 5% is associated with the waveform areas while 30% is used as the peak and time position thresholds. If the comparison of the reference and CUT waveforms results in a value greater than the thresholds for any of these parameters, then the gate transient is considered defective.

Power grid extraction and characterization is done only once for the chip. Generation of test patterns to sensitize the desired paths can be done with available commercial tools (or simple extensions to them). The only steps that are computationally involved are the calibration process and temporal and spatial decomposition. The latter requires solving multiple linear equations for each time sample of the waveforms and can be time consuming. However, the amount of time required for the whole process is orders of magnitude smaller than the normal reverse engineering process that is pursued in failure analysis labs for locating defects.

5.0 Simulation Setup and Results

The simulation experiments performed are designed to provide proof-of-concept. To this end, we use an inverter chain to represent the sensitized paths in our simulations. The layout of the

inverter chain is shown in Figure 11. The ten inverters that define the chain are shown along the bottom of the figure. The upper portion shows inverters that comprise the fan-out that attaches to nodes along the chain[1]. The layout are implemented using the TSMC 0.25 um technology design rules [25].

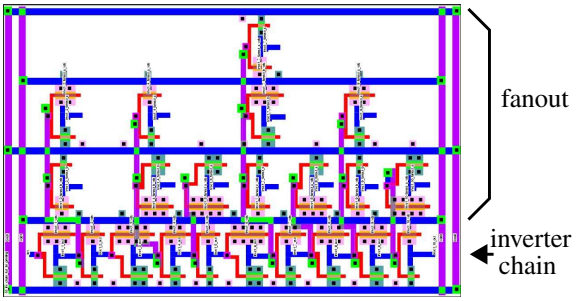

Fig. 11. 10 inverter chain

The layout of the two metal layer power grid used in the simulations is shown in Figure 12. The V_{DD} and GND grid are interleaved and routed in a mesh configuration. Labels are added (not shown) to the bottom metal layer for the attachment of current sources. Labels are also added to the top metal layer to represent the attachment points for the four C4s. The dimensions of the power grid are 500 um X 500 um.

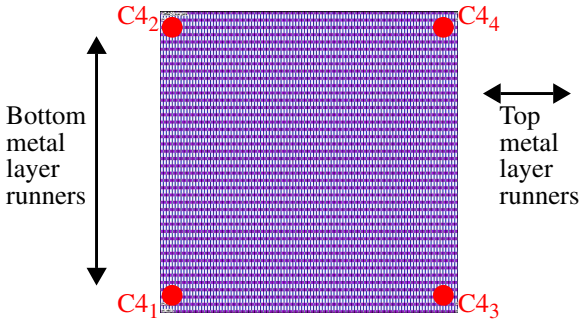

Fig. 12. 2-layer power grid.

Process variations are introduced into the simulation models using the published parameters for the TSMC 0.25 um technology [25]. Eight simulation models of the inverter chain and power

1. Fan-out was added to improve the accuracy of the modeling of an actual circuit path.

grid are extracted using SPACE [26] and a custom extractor, respectively. SPICE is used to simulate the inverter chain and the gate transients through each of the p-channel transistor sources are saved as piece-wise linear waveforms. SPICE is also used to compute the IR functions for the power grid RC models.

One of the process models is designed as the reference and the CUTs are emulated using the remaining seven process models. The CUT's core logic is emulated using ten different configurations of four randomly placed inverter chain paths; three that are defect-free and one path with a defect inserted. The remaining portions of the core logic are modeled using discrete capacitors, which are distributed across the two dimensional layout of power and ground grids. The ten inverter chain configurations are replicated in separate models for each of the nine defects and seven process models for a total of 630 simulation models.

Calibration tests are performed on the RC models of the power grid by attaching a current source to the bottom metal layer at each of the (x, y) positions representing the positions of the calibration circuits (CCs). For the reference diagnostic experiments, current sources are again attached to the bottom layer of the power grid at attachment points given by one of the inverter chain configurations. The IR functions are computed for each of these tests using a current step input.

The defective CUTs are simulated using current sources configured to drive the piece-wise linear waveforms of the inverter chain gate transients onto the attachment points. For each defect experiment, one of the inverter chains is randomly selected and the current source stimuli is replaced with those produced under a simulation of a defective inverter chain. These power grid simulations are repeated under various defect types across the set of process models.

The remaining portion of the diagnostic process is identical to that described in the previous

sections. For example, the reference inverter chain gate transients are calibrated for linear and non-linear process variation effects and spatial and temporal decomposition processes are applied. The CUT's gate transients are then compared with those generated from the inverter chain reference experiments and a decision made with regard to whether an anomaly is detected or not. The diagnostic portion of the simulation experiment flow is shown in Figure 13.

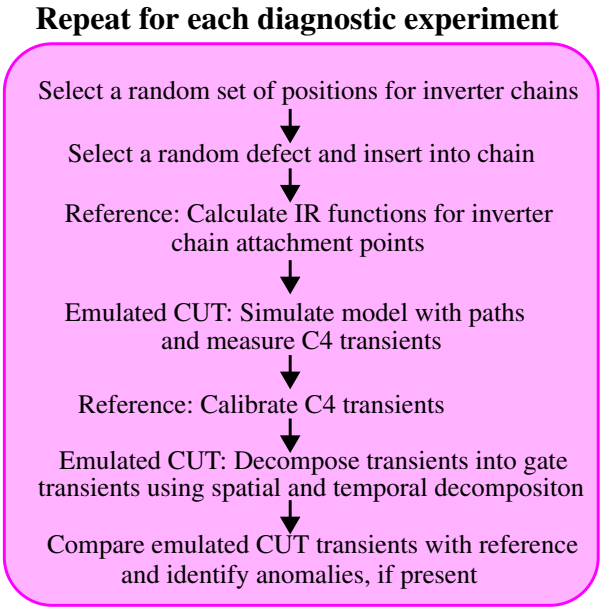

Fig. 13. Flow of simulations

5.1 Defect Types

The defects selected for investigation represent those that are common in today's technologies, and include resistive stuck-at, bridging and open defects. The emulation of defects is accomplished by inserting, removing and modifying resistors that are extracted from the layout shown in Figure 11. The defect types modeled in our experiments include 1) resistive shorts to ground, 2) resistive opens between inverters of the chain and 3) resistive bridges between the outputs of two inverters of the chain. Figure 14 depicts the positions of the inserted defects in a schematic of an inverter chain. Three different resistances for each defect type are investigated. Table 1 lists the

attributes of the nine defect types.

Defect	Type	Value	Position in chain
OP1,2,3	Open	5, 20, 50 kOhms	inv #4
RB1,2,3	Resistive Bridge	3, 6, 10 kOhms	inv #4, #8
RS1,2,3	Resistive Short	1, 9, 50 kOhms	inv #6

Table 1: Defect types used in experiments

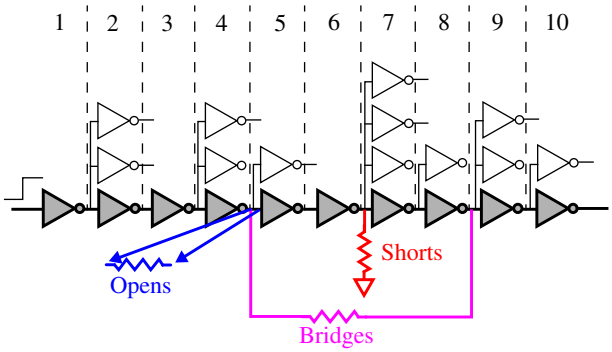

Fig. 14. Inverter chain with fan-out, and positions of the simulated defects.

5.2 Simulation Results

The effectiveness of the spatial and temporal decomposition process (STDP) is evaluated based on three metrics; the number of *correct diagnosis*, the number of *mis-diagnosis* and a third parameter called *uncertainty*. Correct diagnosis is the ratio of the number of times that STDP correctly identifies the defective gate for a particular defect type to the total number of times that defect type is inserted into the paths. Incorrect diagnosis of defect-free gates is accounted for in the mis-diagnosis metric. Mis-diagnosis is computed as the ratio of the number of defect-free gates identified as defective to the total number of gates in the experiment for a specific defect type. Therefore, the overall level of accuracy associated with STDP requires that both the correct diagnosis and mis-diagnosis be considered.

As an example, for cases where no calibration is applied or the applied calibration method is not effective, the correct diagnosis metric will be high because process variations effects will add

to the anomaly produced by the defect. In other words, the gate transient calculated for the defective gate will be different from that produced by the defect-free gate because of both process variations and the defect. However, the false diagnosis metric will also be high under these conditions because the un-calibrated process variation effects will also cause many defect-free gates to be identified as defective.

In order to express both of these characteristics, an *uncertainty* metric is defined as given by Equation 21. The uncertainty metric associated with the calibration techniques is normalized between 0.0 and 1.0 by dividing by the uncertainty value computed when none of the calibration methods are applied. Therefore, uncertainty values near 0.0 indicate good diagnostic results while those near 1.0 indicate poor results.

$$\text{Uncertainty} = \text{false diagnosis} + (1 - \text{success rate}) \times \text{false diagnosis} \quad \textbf{(21)}$$

$$\text{Normalized Uncertainty} = \frac{\text{Uncertainty}}{\text{Uncertainty}_{\text{no calibration}}}$$

5.3 Calibrating Linear Variations Only

The first set of experiments is designed to determine the effectiveness of each of the power grid calibration methods described in Section 3.3.1. Four CCs are used as shown in Figure 12 for the *Matrix-Based Approach*. The *Distributed Calibration Circuit Approach* is implemented by uniformly distributing 16 CCs in one scenario and 30 CCs in a second scenario across the 2-dimensional plane of the power grid (see Figure 3(b)).

In order to evaluate the effectiveness of the linear calibration methods, the gate transients under each of the different process models are used directly in the STPD decomposition process. In other words, non-linear variations are not present in this analysis.

The plots of the results include two base cases that are used to help determine the level of effectiveness of the three calibration techniques. In the first base case, the same process model is

used for the reference and CUT. Therefore, the power grids are identical and calibration is not needed. These results represent the best that can be achieved using STDP. For the second base case, the power grids vary according to the process parameters used to extract them, but no calibration is performed. These results represent the worst case. The results obtained under the three calibration schemes necessarily fall between these two extremes.

Figure 15 shows the correct diagnosis metric as a bar graph. A set of five bars is given for each defect type, one for the base case 'no calibration' (left-most bar), one for each of the IR calibration techniques (middle bars) and one for the second base case 'no process variation' (right-most). Each bar represents the average of seventy individual experiments, counted as ten inverter chain configurations over the seven process models. Figure 16 shows the corresponding mis-diagnosis results. In general, the high values for correct diagnosis and the low values of mis-diagnosis for the bars corresponding to the calibration methods in these figures indicate that the method is effective at diagnosing defects to a specific gate in the inverter chains.

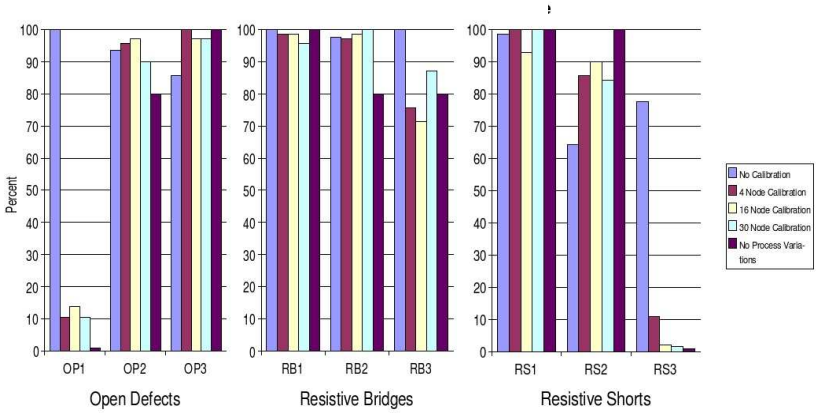

Fig. 15. Correct diagnosis using linear PE calibration only.

Several specific features of these results merit further elaboration. The first notable feature indicates that STDP has some inherent inaccuracies. This is illustrated by the non-zero mis-diagnosis values for OP2, OP3 and RS1 under the ideal 'no process variation' case (right-most bar) in

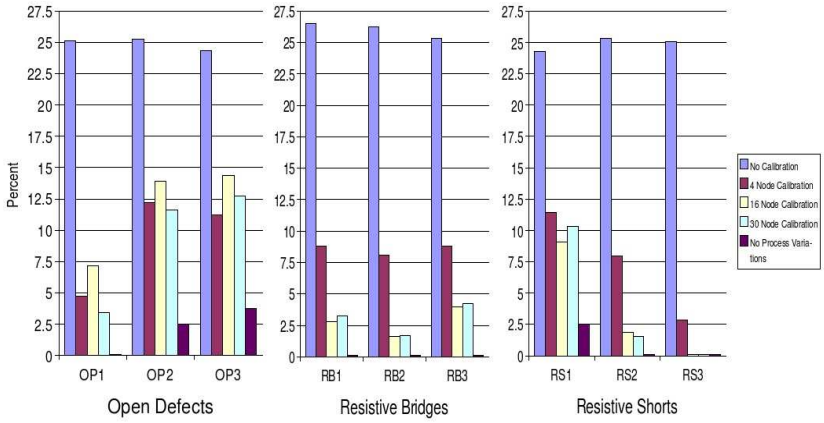

Fig. 16. Mis-diagnosis using linear PE calibration only.

Figure 16. In these cases, STDP incorrectly diagnosis some gates as defective. Since the reference and emulated CUT are identical in these experiments, i.e., they posses no linear or non-linear process variations, the expected result for mis-diagnosis is zero.

A second feature indicates that some defects are more difficult to diagnose than others. This is expected since the magnitude of the transient anomaly introduced by the defect depends on its resistance. For example, the first open defect, OP1, is implemented by adding a 5 kOhm resistor to the connection between two consecutive inverters as shown in Figure 14. This produces only a small amount of additional delay (over the defect-free case) along the inverter chain. The same is true for RS3, which is implemented as a 50 kOhm resistive short. The difficulty of detecting these "subtle" defects is shown by the 'no process variation' bars in the correct diagnosis bar graph of Figure 15, which shows a value of zero for both cases. For other open and shorting defect experiments, the larger magnitude of the anomaly increases the correct diagnosis values in Figure 15 under any of the calibration methods. For bridges, the smaller difference in the magnitude of the transient anomaly introduced by the defect makes this trend less noticeable.

A third feature is that the correct diagnosis for OP1 in Figure 15 is 100% under the 'no calibration' case. This is the most difficult open defect to detect and therefore 0% is the expected value.

Since process variations are not calibrated for in these experiments, STDP identifies most gates (defective and defect-free) as defective. The mis-diagnosis in Figure 16 confirms this analysis, by also showing large values for the 'no calibration' case. A more important result is shown for the calibration techniques, which indicate that these techniques are significantly reducing process variation effects, e.g., their values are within 15% of the ideal 'no process variation' case for open and resistive shorting defects and less than 5% of the ideal 'no process variation' case for bridging faults.

For some defects, there are several instances in Figure 15 where the 'no process variation' bar is smaller than the calibration cases. This is true for the same reason given above, that for some defects and process model combinations, process variations have not been completely corrected by the calibration techniques and the STDP accidentally diagnosis the correct (defective) gate. However, in these cases, the false diagnosis rates are also higher, as given in Figure 16, which reflects the fact that some of the successful diagnosis results are due to process variations in the power grid.

Figure 17 presents a graph of the normalized uncertainty results obtained using Equation 21. The format is the same as that given for Figures 15 and 16. As indicated above, the 'no calibration' scenario produces a value of 1.0 for all defects and the 'no process variation' scenario yields the best results. For open defects, the effectiveness of the calibration techniques are about the same while for bridges and shorts, the distributed calibration schemes out-perform the matrix scheme.

5.4 Calibrating Linear and Non-linear Variations

In this section, the simulation models include both linear and non-linear PE variations and the two non-linear calibration techniques described in Section 3.3.2, the *Neighboring Path* approach

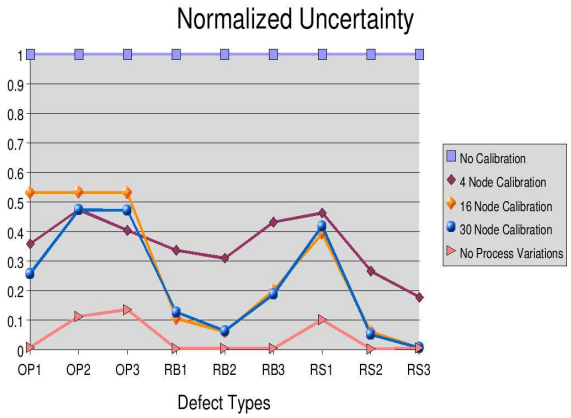

Fig. 17. Normalized uncertainty using linear PE calibration only.

and the *Scaling* approach, are applied.

The *Neighboring Path* approach calibrates for linear and non-linear variations simultaneously. Therefore, we do not apply any of the linear calibration methods in these experiments. The neighboring paths are implemented by placing a copy of a defect-free inverter chain "close" to the defective inverter chain. The defect-free version and the defective version of the inverter chain are extracted using the same process parameters. Averaging as given by Equation 8 is implemented by placing the defect-free inverter chain at three different positions and repeating the experiment.

For the *Scaling* approach, the scaling factors are computed using the corresponding C4 transients from the reference and CUT and the CUT transients are calibrated for non-linear variations. Linear variation is then calibrated for using the three techniques described in the previous section. The results are reported separately for each of the linear calibration approaches. Figures 18 and 19 give the correct diagnosis and mis-diagnosis values under the two base cases, the *Neighboring Path* approach and the three *Scaling* approach scenarios.

From Figure 19, the mis-diagnosis values increase over those shown in Figure 16 with the inclusion of non-linear variations, as expected. Also, the trend in the mis-diagnosis values across the linear calibration methods is similar to those shown in Figure 16, i.e., mis-diagnosis increases

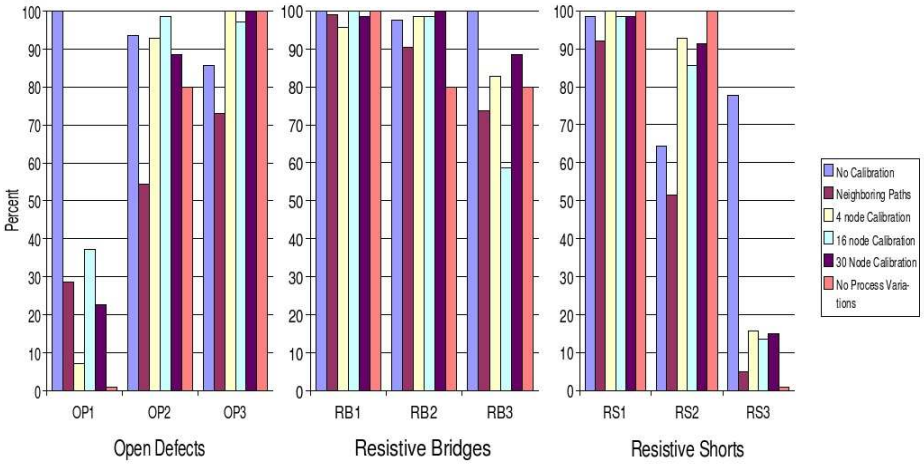

Fig. 18. Correct diagnosis using linear and non-linear PE calibration.

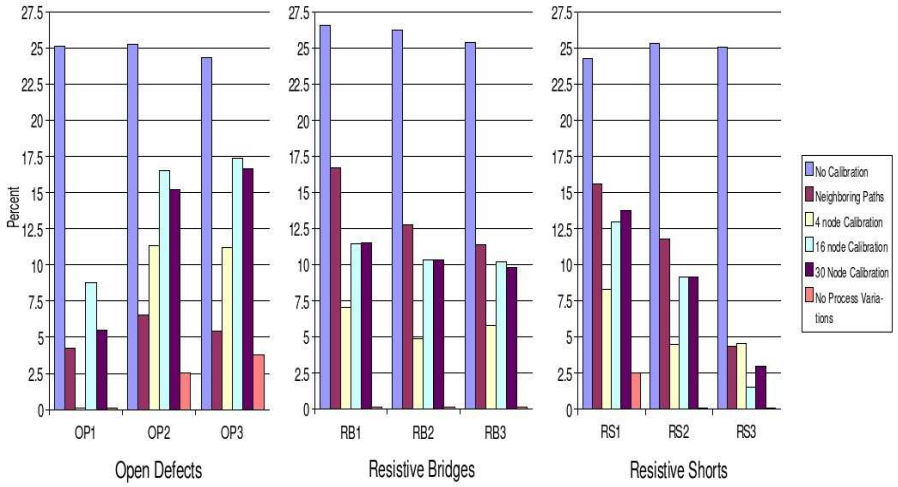

Fig. 19. Mis-diagnosis using linear and non-linear PE calibration.

as the defect's transient anomaly increases for opens and shorts. This occurs because the larger transient anomaly 'leaks' into surrounding defect-free gate transients during the decomposition process, making it more difficult to isolate the anomaly to a specific gate.

It is apparent from Figures 18 and 19 that the Neighboring Path approach performs fairly well in comparison with the three Scaling approaches, where for example, the mis-diagnosis values are actually smaller for open defects and only slightly larger for resistive bridges and shorts. The

advantage of using the *Neighboring Path* approach is that it does not require any additional hardware. Therefore, it can be applied to existing CUTs as long as it is possible to independently sensitize paths close to the defect candidate paths.

The normalized uncertainty values are given in Figure 20. The line associated with '4 node calibration', which reports the results for Matrix calibration (for linear variations) with Scaling (for non-linear variations), provides a lower uncertainty for most defect types, and is therefore the best approach overall. Although this approach requires support hardware, e.g., calibration circuits, the overhead is small, particularly when compared with the 16 node and 30 node calibration schemes.

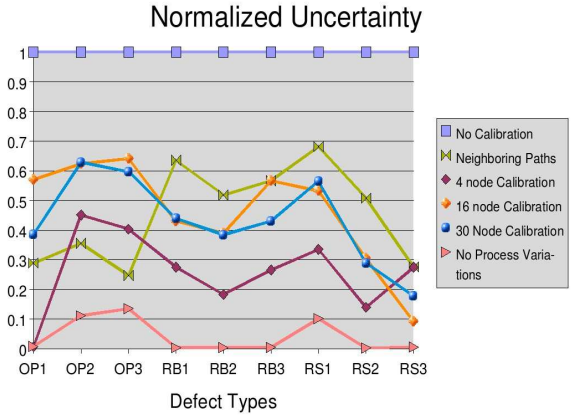

Fig. 20. Normalized uncertainty using linear and non-linear PE calibration.

6.0 Conclusions

A novel diagnostic method is proposed that is based on the analysis of the gate transients produced as signal propagate along paths in the core logic. The gate transients are estimated from the measured power port (C4) transients using a temporal and spatial decomposition process. The presence of a defect will change the shape of the transient signal generated by a gate. By comparing the gate transients derived from decomposing the CUTs C4 transients with those obtained from defect-free simulation experiments, it is possible to identify the anomaly and localize the defect.

Process and environmental (PE) variations are significant in advanced technologies and adversely impact the diagnostic resolution of our proposed method. Three linear and two non-linear PE calibration techniques are proposed and investigated using simulation experiments. The linear calibration methods reduce transient signal variations caused by impedance variations in the power grid and test environment while non-linear calibration methods deal with process variations that affect delay characteristics. Each calibration method is able to improve diagnostic resolution over the base case of no calibration.

Simulation experiments are performed to demonstrate the effectiveness of the method on a variety of defect types. In many cases, the defective gate is successfully identified and the overall diagnostic resolution allows the position of the defect to be confined to a single gate or a small number of gates.

7.0 References

[1] ITRS (http://public.itrs.net/)

[2] Chakravarty S., Liu M., "IDDQ Measurement Based Diagnosis of Bridging Faults", *JETTA*, Vol. 3, pp. 377-385, 1992.

[3] Gong Y. and on Chakravarty S., "On adaptive Diagnostic Test Generation, *ICCAD*, pp. 181-184, 1995.

[4] Yiming G., Chakravarty S., "Using Fault Sampling to Compute IDDQ Diagnostic Test Sets", *VTS*, pp. 74-79, 1997.

[5] Pomeranz I. and Reddy S. M., "On the Generation of Small Dictionaries for Fault Location", *ICCAD*, 1992, pp. 272-279.

[6] Aitken R. C., "A Comparison of Defect Models for Fault Location with IDDQ Measurements", *ITC*, 1992, pp. 778-787.

[7] Naik S., Maly W., "Computer-Aided Failure Analysis of VLSI Circuits using IDDQ Testing", *VLSI Test Symposium*, 1993, pp. 106-108.

[8] Millman S. D., Acken J. M., "Diagnosing CMOS Bridging Faults with Stuck-At, IDDQ, and Voting Model Fault Dictionaries", *Custom Integrated Circuits Conference*, 1994, pp. 409-412.

[9] Isern E., Figueras J., "IDDQ Test and Diagnosis of CMOS Circuits", *Design & Test of Computers*,

Volume 12, Issue 4, 1995, pp. 60-67.

[10] Young-Jun K., Walker D. M. H., "Yield Learning via Functional Test Data", *ITC*, 1995, pp. 626-635.

[11] Sanada M., "A CAD-based Approach to Failure Diagnosis of CMOS LSI's using Abnormal Iddq", *VLSI Test Symposium*, 1996, pp. 186-191.

[12] Nigh P., Forlenza D., Motika F., "Application and Analysis of IDDQ Diagnostic Software", *ITC*, 1997, pp. 319-327.

[13] Gattiker A. E., Maly W., "Current Signatures: Application", *ITC,* 1997, pp. 156-165.

[14] Thibeault C., Boisvert L., "Diagnosis Method Based on Delta IDDQ Probabilistic Signatures: Experimental Results", *ITC*, 1998, pp. 1019-1026.

[15] Lavo D. B., Larrabee T., Colburn J. E., "Eliminating the Ouija(R) Board: Automatic Thresholds and Probabilistic IDDQ Diagnosis", *ITC*, 1999, pp. 1065-1072.

[16] Acharyya D., Plusquellic J., "Calibrating Power Supply Signal Measurements for Process and Probe Card Variations", *Workshop on Current and Defect Based Testing*, 2004. pp. 23-30.

[17] Plusquellic J., Acharyya D., Tehranipoor M. and Patel C., "Triangulating to a Defect's Physical Coordinates Using Multiple Supply Pad IDDQs: Test Chip Results", *International Symposium on Testing and Failure Analysis*, Nov. 2006, pp. 36-45.

[18] Plusquellic J., Acharyya D., Singh A., Tehranipoor M. and Patel C., "Quiescent-Signal Analysis: A Multiple Supply Pad IDDQ Method", *Design and Test of Computers*, Volume 23, Issue 4, April 2006, pp. 278-293.

[19] Muhammad K., Roy K., "Fault Detection and Location using IDD Waveform Analysis", *Design and Test of Computers*, Volume 18, Number 1, pp. 42-49, 2001.

[20] de Paul I., Rosales M., Alorda B., Segura J., Hawkins C., Soden J., "Defect Oriented Fault Diagnosis for Semiconductor Memories Using Charge Analysis, Theory and Experiments", *VLSI Test Symposium*, 2001, pp. 286-291.

[21] Bhunia S., Roy K., Segura J., "A Novel Wavelet Transform Based Transient Current Analysis for Fault Detection and Localization", *DAC*, 2002, pp. 361-366.

[22] Singh A., Patel C. and Plusquellic J., "Fault Simulation Model for iDDT Testing: An Investigation", *VLSI Test Symposium*, pp. 304-310, 2004.

[23] Tech report, available on request.

[24] Wiener N., "Extrapolation, Interpolation, and Smoothing of Stationary Time Series", New York: Wiley, 1949.

[25] http://www.mosis.com/Technical/Testdata/tsmc-025-prm.html

[26] van Genderen A., van der Meijs N., Beeftink F. , Elias P. , Geigenmuller U. and Smedes T., "SPACE,

Layout to Circuit Extraction software module of the Nelsis IC Design System", Delft University of Technology, 1996. (space@cas.et.tudelft.nl).