Impedance Profile of a Commercial Power Grid and Test System

Dhruva Acharyya and Jim Plusquellic

Department of CSEE, University of Maryland, Baltimore County

Abstract

An impedance profile of a commercial power grid and a tester power distribution system is developed in this paper. The profile is used to identify the measurable frequency range of the power supply transient signals generated by a chip. Several resistance-capacitance (RC) models of the power grid are analyzed to determine the impact of each capacitance type. The impedance profile of a C4-based production testing environment is then developed. The impedance profile of the combined probe card and the power grid RC models illustrates the range of frequencies that are measurable at the supply ports of the chip-under-test (CUT). The results suggest that it is possible to measure the important frequency components of a chip's power supply transients in a production test environment for use in fault detection and localization procedures.

1.0 Introduction

Conventional testing methods are challenged by changing circuit sensitivities and emerging defect mechanisms resulting from the use of new fabrication materials in very deep submicron processes [1]. For example, the change from a subtractive aluminum process to damascene Cu may lead to more particle-related blocked-etch resistive opens. Technology scaling also increases the probability of resistive vias caused by incomplete etch. The additional delays introduced by these types of resistive defects in combination with increased circuit sensitivity due to shorter clock cycles, reduced timing slack, crosstalk and PWR/GND bounce increase the likelihood of random defects causing delay fails.

Similarly, 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. 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.

We believe that power supply testing methods are well aligned with these needs and others as described in the International Technology Roadmap for Semiconductors. In our previous work, a testing method is presented for fault detection that uses correlation analysis of multiple simultaneously-measured power supply transient signals [2]. The transients at each of the supply ports of a chip-under-test (CUT) are cross-correlated to reduce the adverse effects of process variations on fault detection resolution. The multiple supply port measurements are analyzed for the regional signal anomalies introduced by defects. The regression analysis technique that we propose in [3] is able to detect anomalies in the ratios of the waveform areas of signals measured at different topological locations on a defective chip.

This type of testing strategy has several advantages. The use of multiple individual supply port measurements suggests that fault detection resolution will not degrade as rapidly as other single point measurement techniques as chips size and transistor density increase. Secondly, aberrations in the individual power port signals introduced by defects can be used to estimate the position of the defect in the physical layout of the chip.

However, in order for power supply transient methods to be useful, it must be possible (and practical) to measure these signals with sufficient frequency resolution. The uncertainty in measurement resolution derives from the filtering characteristics of the resistive, capacitive and inductive elements of the power grid and the power distribution system (PDS) of the tester and probe card. The impact of transistor source, N-well and on-chip decoupling capacitance are of particular interest since many have expressed concern over the fault detection sensitivity of power supply analysis methods, particularly if the successful detection of faults depends on the ability to measure the high-frequency signal content of the transients.

This paper explores the characteristics of the RC model of the supply grid and the RLC model of the PDS as a means of validating power supply transient testing methods. Our analysis indicates that it is possible to measure frequencies up through 1 GHz without significant attenuation at measurement points close to the CUT. Such access is possible at wafer probe using, for example, a modified membrane probe card designed to allow measurements at each of the power supply ports of the CUT.

This paper is organized as follows. Section 2.0 gives a short background on other proposed transient current (I_{DDT}) methods. Section 3.0 describes the details of the commercial power grid used in the simulation experiments.

This work is supported by a Faculty Partnership Award from IBM's Austin Center for Advanced Studies (ACAS) Program and by an NSF grant, award number 0098300.

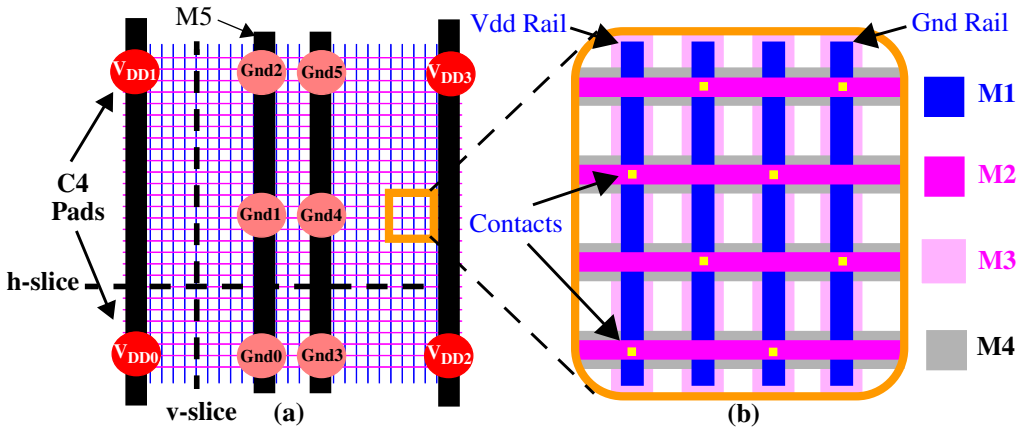

Figure 1. The "Quad": A portion of the commercial power grid used in the simulation experiments.

Section 4.0 discusses the various capacitance models of the power grid that were extracted and analyzed. Section 5.0 gives a model for a power distribution system (PDS) and Section 6.0 illustrates the effects of the PDS on the impedance characteristics of the grid. Section 7.0 provides a brief conclusion.

2.0 Background

There are a variety of testing and fault localization methods proposed that are based on the analysis of power supply signals [4-10]. The main drawback of the techniques proposed in [4-7] is that they do not account for vector-to-vector or process variations. Therefore, they are difficult to apply to devices fabricated in advanced technologies, in which these types of variations are significant and must be accounted for. The ECR I_{DDT} method accounts for process variation effects by computing ratios from the time domain I_{DDT} waveform areas measured under different test sequences [8]. However, the analysis presented in this paper suggests that methods based on a single test point measurement made between the probe card and the tester's power supply will not provide adequate resolution to enable the detection of resistive defects such as those described above. Similar concerns hold true for the fault localization methods proposed in [9] and [10].

The detection and diagnostic methods that we propose make use of regional signal information available in the individual supply port measurements. A major issue of our method (and others) is the ability to reliably measure frequency components of the transient signals that are key to the detection and localization of faults. Therefore, the objective of this work is to determine the frequency content of signals measured at various access points on the tester and probe card. Although the focus of this research is on signal measurement points close to the CUT, a comparative analysis is presented using the signals measured on the PWR plane of the probe card to illustrate the effect on resolution for single test point measurement methods.

3.0 Simulation model

Figure 1(a) shows a portion of the commercial power grid under analysis in this work, that is subsequently referred to as the Quad. The Quad occupies a 10,000 by 10,000 unit area and interfaces to a set of external power supplies through an area array of V_{DD} and GND C4 pads. A C4 pad is a solder bump for an area array I/O scheme. As indicated in Figure 1(a), there are four V_{DD} C4s and 6 GND C4s in this portion of the grid.

Figure 1(b) expands a portion of the grid and shows that it is constructed over 4 layers of metal, with metal 1 (M1) and metal 3 (M3) running vertically and M2 and M4 running horizontally. The C4s are connected to wide runners of vertical M5, shown in Figure 1(a), that are in turn connected to the M1-M4 grid. In each layer of metal, the V_{DD} and GND rails alternate. The alternating vertical V_{DD} and GND rails are connected together using alternating horizontal metal runners. Stacked contacts are placed at the appropriate crossings of the horizontal and vertical rails.

4.0 Capacitance models of the power grid

We derive several RC models of the Quad using an extraction script that preserves the physical structure of the metal interconnect in the topology of the RC network, i.e. no network reduction heuristics are applied. The resistance per square and the overlap capacitances per unit area of TSMC's 0.25 µm 5 metal process used in the extraction process were obtained from published parameters [11]. The focus of this analysis is on the capacitance characteristics of the grid. The resistance model is presented but is described in more detail in [12].

One of our key interests is to determine the major capacitance components of the grid that are responsible for attenuating the core-logic-generated transients at the C4s (the measurement points). Our analysis reveals that there are five distinct capacitance types. The left side of Figure 2 shows two wire-related capacitances labeled C_{self} and

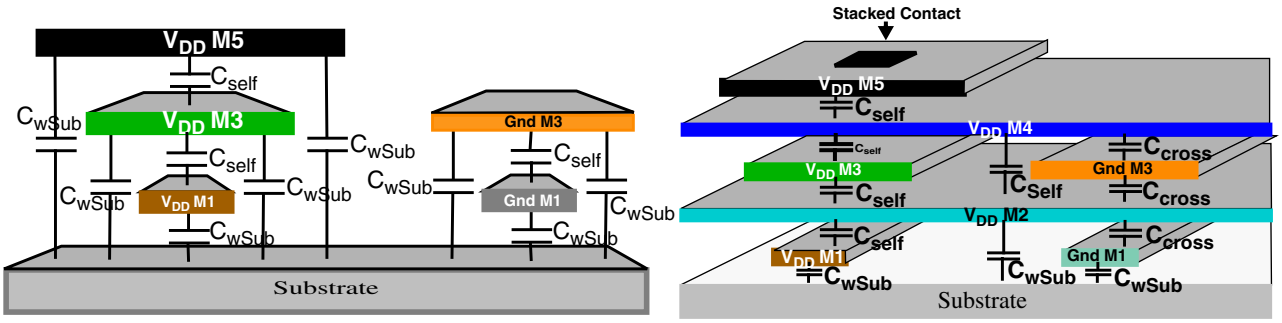

Figure 2. Self, substrate and cross wire-related capacitances of the Quad.

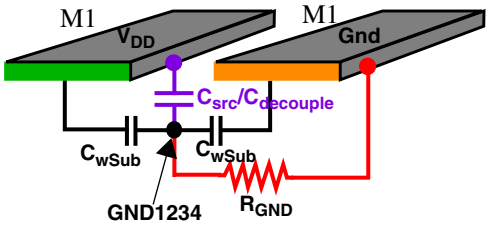

Figure 3. Source and decoupling cap. of the Quad.

C_{wSub} in a cross-section of the power grid shown in Figure 1(b). C_{self} is capacitance between two V_{DD} or two GND metal runners in different layers. The large overlap region of these runners, as given by the arrangement of the grid, reduces the impedance between points in the substrate and the C4s. However, two thick oxide layers separate these runners, which reduce the significance of the coupling effect. For example, the M1-M2 capacitance per unit area is 39 aF/um^2 and drops to 15 aF/um^2 for M1-M3 (see MOSIS model *n99y* [11]). C_{wSub} is the overlap capacitance of the exposed portions of the wires to the substrate. The widening of the metal runners from M1 to M5 introduces coupling between all metal layers and the substrate, although the effect is small for upper layers. For example, M5-substrate capacitance per unit area is only 8 aF/um^2. The third wire-related source of capacitance is shown on the right side of Figure 2 labeled C_{cross}. This is the inter-layer capacitance between V_{DD} and GND. The configuration of the grid suggests that the cross coupling effect is small since the V_{DD}-GND overlap regions are minimized in this style of grid design.

The last two sources of capacitances are shown in Figure 3 labeled as $C_{src}/C_{decouple}$. C_{src} represents the P-diffusion source and N-well capacitance while $C_{decouple}$ represents the inserted on-chip decoupling capacitance. In our power grid model, we distribute both of these capacitances across approximately 3,000 evenly distributed points in M1 on the V_{DD} grid. The ground node of these capacitors is represented by nodes of the form GND*xxx*, as shown in the figure for node GND1234. These nodes also

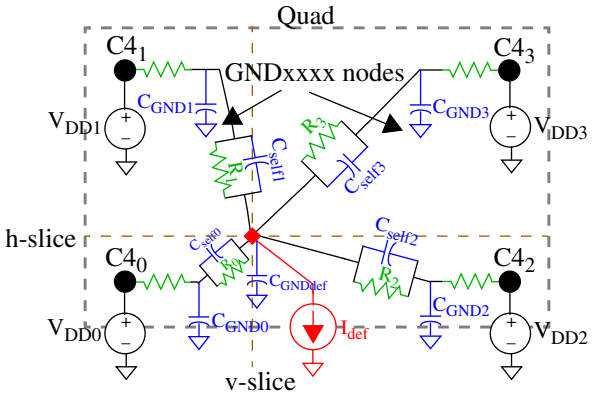

Figure 4. Lumped RC model of the Quad.

provide the ground connections for the wire-generated substrate capacitances, C_{wSub}, located in the region of the $C_{src}/C_{decouple}$ node. Each of the GND*xxxx* nodes connect back up to the GND grid through 1 Ω resistors, R_{GND}. These resistors are designed to model the equivalent series resistance of the on-chip decoupling capacitors. The values of the inserted source and decoupling capacitance are representative of those present in commercial designs.

We extracted a set of RC models from the Quad layout, each derived using one or more of the capacitance types identified above. The total contribution of C_{self}, C_{wSub} and C_{cross} derived from the models was determined to be less than 2% when expressed as a percentage of total capacitance. As the following analysis demonstrates, the impact of these wire-related capacitances on the impedance profile of the grid is small when compared to the impact of C_{src} and $C_{decouple}$. These latter capacitances function to further isolate the higher frequency components of the transient signals to smaller regions of the grid over that provided by resistance alone.

In order to evaluate the impact of each of these capacitances, a set of simulations were performed with a current source placed between the V_{DD} and GND M1 runners at the position represented by the crossing of lines labeled 'v-slice' and 'h-slice' in Figure 1(a). Figure 4 shows an equivalent RC model of the Quad illustrating the dominant

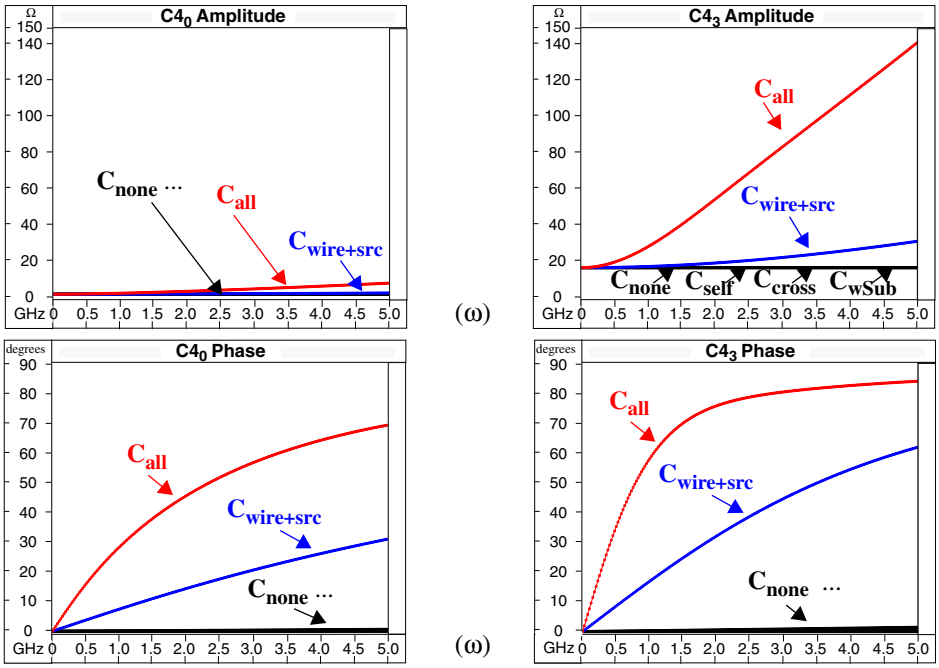

Figure 5. Impedance characteristics at $C4_0$ and $C4_3$ with an 'ideal' power supply computed using an AC sweep of a current source tied to M1 at intersection of v-slice and h-slice lines in Figure 1(a).

current paths between the current source labeled I_{def} (for defect) and $C4_0$ through $C4_3$. Here, C_{GND} represents the equivalent capacitance derived from one or more of C_{cross}, C_{wSub}, C_{src} and $C_{decouple}$ (depending on the model). The probe card model is not included in this analysis since we want to determine the impedance characteristics of the grid alone. The results obtained from this analysis represent the 'ideal' measurement environment. These are used as a reference for the analysis presented in Section 6.0 that includes a model of the off-chip power distribution system.

A set of AC simulations were performed in SPICE over the frequency range from 1 Hz to 5 GHz in 25 MHz steps with the current source configured as an AC source. This type of analysis allows the impedance characteristics of the grid to be computed from the current source to each of the C4s under each of the simulation models. Since the current source is closer to $C4_0$ than to the other C4s, its impedance is expected to be least effected by the low-pass filter properties of the grid's RC network.

Equation 1 was used to process the SPICE AC analysis data into "equivalent impedance" between the C4s and the current source. Here, $V_{def}(\omega)$ is the voltage vector on the

$$Z_{eq}(\omega) = \frac{V_{def}(\omega)}{I_{VDDx}(\omega)} \qquad (1)$$

positive terminal of the current source and $I_{VDDx}(\omega)$ is the current vector through each of the four voltage sources x (labeled V_{DD0} through V_{DD3} in Figure 4). The division operation is carried out by dividing the amplitude (resistance) components and subtracting of the phase components. Figure 5 plots the equivalent impedance, Z_{eq}, at $C4_0$ on the left and at $C4_3$ on the right, as a function of frequency, ω, from 1 Hz up through 5 GHz. The amplitude plots are shown above the phase plots.

Each plot contains six superimposed waveforms, one from each of the capacitance models, C_{none}, C_{self}, C_{wSub}, C_{cross}, $C_{wire+src}$, and C_{all}. Only the $C4_3$ amplitude plot contains labels for all six models. C_{none} is the resistance only model, C_{self}, C_{wSub} and C_{cross} are the wire-related capacitance models, $C_{wire+src}$ includes the three wire-related capacitances + C_{src}, and C_{all} adds $C_{decouple}$ to the $C_{wire+src}$ model.

The impedance at both $C4_0$ and $C4_3$ under the C_{none} model is resistive only, i.e. resistance is constant and phase is 0 at all frequencies. The difference in the resistance at DC between $C4_0$ and $C4_3$ is significant, e.g. 3.3 Ω vs. 18 Ω. This suggests that the magnitude of the I_{DDQ} measured at $C4_0$ would be a factor of 5 greater than the magnitude measured at $C4_3$ for a defect at this position in the layout (assuming the probe card resistance is small relative to the grid resistances). The strong regional I_{DDQ} behavior created by the resistance characteristics of the power grid can be used as a means of localizing defects [12].

The addition of capacitance to the model increases the impedance amplitude for higher frequencies with respect to

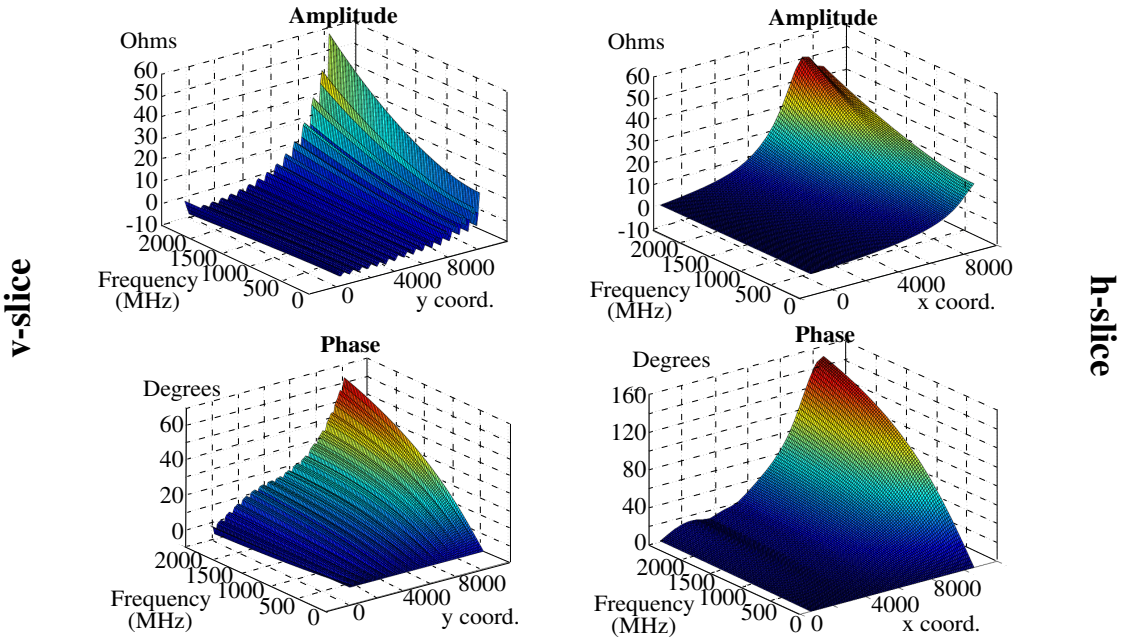

Figure 6. Impedance differences in $C4_0$ along v- and h-slices of Figure 1(a).

the DC resistance, as expected. However, the impact of capacitance on impedance amplitude is strongly dependent on the capacitance elements included in the model. For capacitance models C_{self}, C_{wSub} and C_{cross}, the impact is very small. The curves for these models (indicated by 'C_{none} ...' in the other plots of Figure 5) are nearly indistinguishable from the DC line given by the C_{none} model.

Under the $C_{wire+src}$ and C_{all} capacitance models, the impedance amplitude is a strong function of the distance between the source of the transient and the measurement point. For example, the increase in impedance amplitude at $C4_0$ is small, increasing from 3.3 Ω at DC to approximately 10 Ω at 5 GHz under the C_{all} model. Therefore, only a moderate degree of attenuation is expected at $C4_0$ for transient signals generated in the lower left portion of the Quad. This low impedance path, in combination with the more significantly elevated impedance characteristics of other paths, increases the degree of isolation of the higher frequency transient signal components generated in this portion of the grid over that provided by the DC resistance alone. For example, the increase in the impedance amplitude at $C4_3$ is more significant under the C_{all} model, from approximately 18 Ω at DC up to 150 Ω at 5 GHz.

In previous work, we determined that the frequency band from 300 MHz up to the operational frequency, e.g. 1 GHz, contains most of the useful information for the detection of defects and for characterizing delay in the chip [13]. From the amplitude vs. frequency plots in Figure 5, the increase in impedance amplitude at 1 GHz is small (~4.5 Ω for $C4_0$ and ~30 Ω for $C4_3$), which indicates that it is possible to measure these frequency components without significant attenuation in this 'ideal' measurement environment.

The differences in the phase spectra of the I_{DDT}s at both C4s is more dramatic than the differences in their impedance amplitudes. For example, the values of phase at $C4_0$ increase from 0 at DC to approximately 70 degrees at 5 GHz under the C_{all} model. A slightly more significant non-linear trend is evident in the phase spectrum of $C4_3$, where phase increases to approximately 85 degrees at 5 GHz.

In order to evaluate the impedance of the grid for other positions in the layout, a series of simulations were run with the current source placed at discrete points along the lines labeled 'v-slice' and 'h-slice' in Figure 1(a). Since the objective of this analysis is to determine how the grid affects the core-logic-generated transients as a function of position, the amplitude and phase values are computed relative to the values produced under two reference simulations. Figure 6 shows the amplitude and phase difference values up through 2.5 GHz along the v-slice (left) and h-slice (right) lines in the Quad for signals measured at $C4_0$. The reference locations are the bottom-most and left-most points on the v-slice and h-slice lines, respectively. Equation 2 gives the expression that defines the difference operation.

$$Z_{diff}(\omega) = \frac{V_{\text{def}}(\omega)}{I_{\text{VDD0}}(\omega)} - \frac{V_{\text{def\_ref}}(\omega)}{I_{\text{VDD0\_ref}}(\omega)} \qquad (2)$$

The v-slice amplitude plot (top left plot) shows the

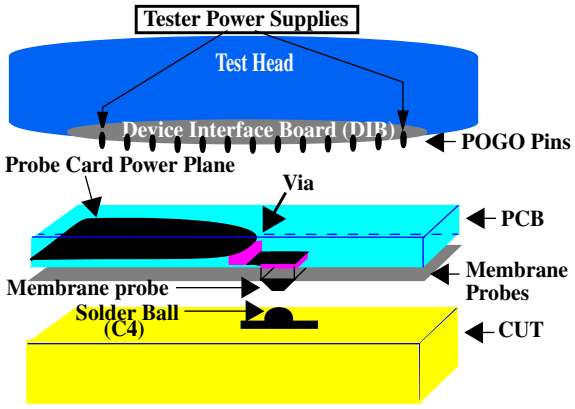

Figure 7. Physical model of power distribution system.

impedance amplitude increasing from 0 Ω at the reference point to 16 Ω at the upper-most point at DC. In contrast, the range at 2.5 GHz is 0 to 58 Ω. (The h-slice amplitude ranges are similar.) The larger amplitude range for the 2.5 GHz component confirms that the grid's capacitive elements restrict the influence of the higher frequency components to smaller regions of the grid in comparison to the lower frequency components. This characteristic may be useful as a means of increasing the resolution of fault detection and localization algorithms that are based on the analysis of signal anomalies introduced by defects in the individual supply port signals. The attenuation characteristics of the grid suggest that the fault introduced by a defect can be more clearly observed as a regional disturbance by analyzing the higher frequency components of the transient signals. Since the amplitudes of the higher frequency components are smaller, the maximum frequency is limited in practice by the signal-to-noise level of the measurement environment.

The phase shift plots reflect the magnitude of the delay introduced by the grid's RC elements. The amount of phase shift is dependent, like amplitude, on the frequency and the position of the current source. Unlike amplitude, there is a large difference in the range of values for the v-slice and h-slice plots, i.e. 0 to 63 degrees vs. 0 to 151 degrees. Moreover, the surface features of the phase plots are unique in comparison to the amplitude surfaces, indicating that phase offers a unique source of (delay) information that may also be useful in testing procedures.

5.0 Probe card model

An important, and often underestimated component of a transient circuit model is the impedance profile of the power distribution system (PDS). The main components of the PDS are the tester's power supply(s), the device interface board (DIB), probe card (PCB) and the membrane probes. Figure 7 shows a simplified physical model of a production test environment with labels corresponding to these elements.

The PDS is responsible for delivering power to the CUT, which sinks current over a wide range of frequencies from DC through several GHz. An upper bound on the frequency range is approximated by Equation 3, where T_r is rise time of the fastest logic signals in the core logic of the CUT [14]. Logic signals propagating in the CUT will

$$F_{knee} = \frac{0.5}{T_r} \quad (3)$$

remain largely un-distorted if the PDS is able to provide a flat frequency response up to this *knee* frequency. For example, the upper limit is approximately 5 GHz for gates with rise times of 100 ps. Note that the impedance of the PDS must remain low for lower frequency current transients as well, such as those that are produced from the propagation of signals along logic paths in the CUT. The tester's power supply alone is capable of providing a flat frequency response only across a very limited frequency range, usually less than a few 10s of KHz, so it is necessary for the PDS to incorporate other elements in order to meet the higher frequency requirements of the CUT.

Another way of analyzing the PDS requirements is to compute a target impedance [15] that the PDS must achieve in order to keep the supply voltage droop under a specified limit. For example, Equation 4 gives the target impedance, Z_{target}, at 6.25 mΩ for a 2.5 V CUT that generates 20 A current transients. In this case, the maximum ripple voltage is specified as 5% of the nominal. Bear in mind that this is

$$Z_{\text{target}} = \frac{(\text{Supply voltage})(\text{Allowed ripple \%})}{\text{current}} \quad (4)$$

$$Z_{\text{target}} = \frac{2.5V \times 0.05}{20A} = 6.25m\Omega$$

the target impedance for the entire CUT. Many CUTs have multiple V_{DD} ports. It follows that the 'per V_{DD} port' target impedance to the tester's power supply can be larger by a factor proportional to the number of ports, as described below.

Figure 8 shows the RLC model of the PDS from the tester power supply through one V_{DD} C4 contact probe on the CUT. (The RL elements of the GND C4 model are also shown but are excluded in this analysis). The tester power supply model is shown on the far left. The series/parallel RL structure of this model and the element values were obtained from the voltage regulator model proposed in [15]. Three of the component values are insensitive to the frequency characteristics of the supplied current. The fourth component (the inductor of value 94 nH) is computed using Equation 5, based on a 2.5 V source, a 5% tolerable variation and an 15 µs response requirement.

$$L_{slew} = V\frac{dt}{di} = (2.5V \times 0.05)\frac{15\mu s}{20A} = 94nH \quad (5)$$

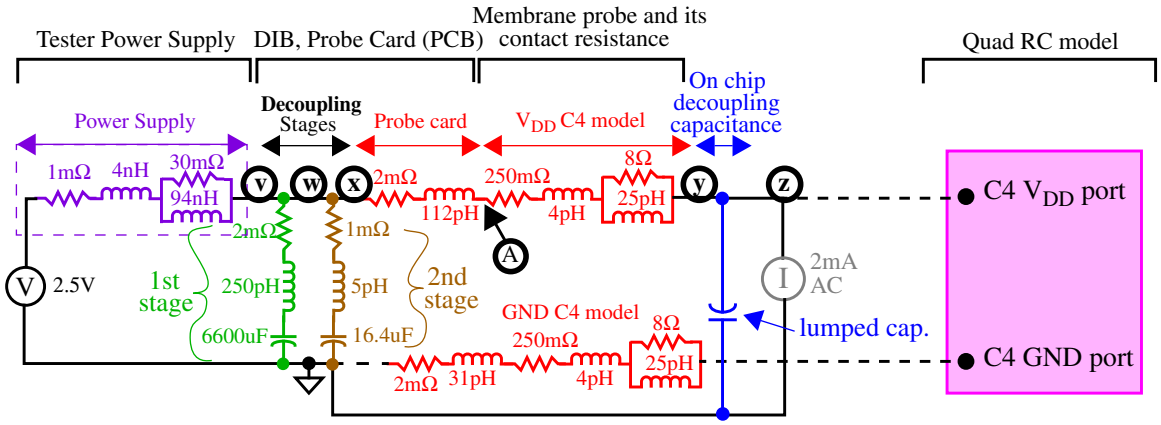

Figure 8. RLC model of physical model shown in Figure 7.

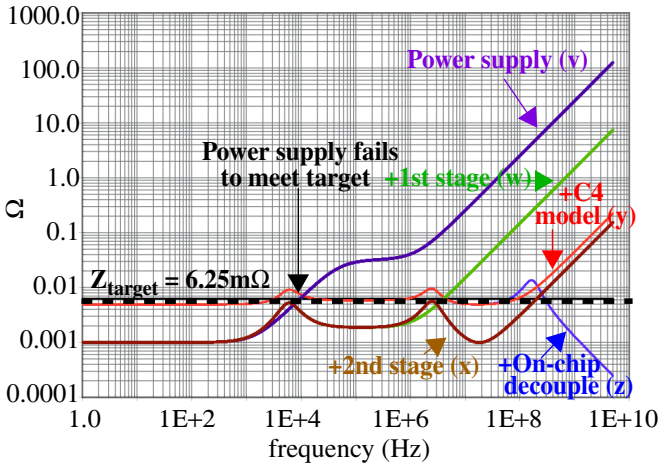

Figure 9. Impedance profile of the PDS as various elements are added, shown left to right in Figure 8.

The output impedance characteristics of the power supply prevent it from responding properly to high-frequency transients generated by the CUT. This can be demonstrated by evaluating the frequency response of the power supply model using the AC analysis feature of SPICE. In order to restrict this analysis to the power supply alone, only the portion of the circuit shown to the left of point *v* in Figure 8 is included in the model. The stimulus consists of a 2 mA current source connected between point *v* and GND. The amplitude (resistance) portion of the impedance is computed by dividing each of the voltage amplitude components at point *v* by 2 mA. The curve is labeled 'Power Supply' in the log-log plot of amplitude vs. frequency in Figure 9. This curve illustrates that the power supply fails to meet the target impedance of 6.25 mΩ (delineated by the dotted horizontal line) at 9 KHz.

The impedance profile of the PDS can be controlled (to some degree) through the proper configuration of the capacitive elements labeled as 'Decoupling stages' in Figure 8. Real capacitors also possess *equivalent series inductance* (ESL) and *equivalent series resistance* (ESR) that must be accounted for. Figure 8 shows these elements in series with the capacitors for each decoupling stage. The ESL of the capacitor limits the highest effective decoupling frequency of the capacitor, while the ESR adversely affects the lower frequency limit, as illustrated in the following analysis.

The first stage of decoupling is designed to maintain the impedance of the PDS below the target at frequencies above the power supply cut-off frequency, e.g. 9 KHz. Equation 6 can be used to determine a lower bound on the amount of capacitance required [14]. A more accurate

$$C = \frac{1}{2\pi f Z_{target}} = \frac{1}{2\pi \times 9\text{E}3 \times 0.00625} = 2829uF \qquad (6)$$

expression that accounts for the ESL and ESR of the capacitors is given by Equation 7. Solving this equation for *C* and

$$Z = \sqrt{R^2 + \left(2\pi fL - \frac{1}{2\pi fC}\right)^2} \qquad (7)$$

solving for *C* yields: $C = \dfrac{1}{2\pi f\left(2\pi fL + \sqrt{Z^2 - R^2}\right)}$

$$C = \frac{1}{(2\pi)9\text{E}3\left((2\pi)9\text{E}3 \times 250pH + \sqrt{6.25\text{E-}3^2 - 2\text{E-}3^2}\right)} = 2979uF$$

setting ESR to 2 mΩ and ESL to 250 pH yields a capacitance of 2979 µF. The selection of the appropriate values for ESR and ESL should be derived from the types of capacitors that are commercially available and from a physical model that allows a realistic estimate of how many capacitors will fit on the probe card.

For example, assume the probe card can accommodate approximately 20 capacitors for the first stage of decoupling. The estimate derived from Equation 7 indicates that at least 2979 uF of capacitance is required. EPCOS manufactures 330 uF ultra low ESR tantalum capacitors with an ESR and ESL of 40 mΩ and approximately 5 nH, respec-

tively. Therefore, a parallel configuration using this type of capacitor yields a total effective ESR of 40 mΩ/20 = 2 mΩ, a total effective ESL of 5 nH/20 = 250 pH and a total capacitance of 20*330 uF = 6600 uF. If the effective ESR of the array is greater than 6.25 mΩ or its total capacitance is less than 2979 uF, then an alternative configuration and/or capacitor type would be required.

The impedance curve for the power supply plus first stage of decoupling is labeled '+1st stage' in Figure 9. This curve was generated by simulating the circuit to the left of point *w* shown in Figure 8 and calculating the result in a manner similar to that described above for the power supply analysis. The effective ESL of 250 pH sets the upper frequency limit of this decoupling stage to approximately 4.0 MHz.

A similar approach is used to design the second stage of decoupling. In this case, a smaller total capacitance is needed but the capacitors must have a smaller effective ESL to be useful. The capacitance of this stage must take over at 4.0 MHz. The total capacitance computed using Equation 7 is 7.5 uF (assuming ESR and ESL values of 3.5 mΩ and 5 pH).

For example, assume the probe card can accommodate 20 second stage capacitors. This constrains the lower limit of each capacitor to 375 nF. AVX corporation manufactures 820 nF ceramic capacitors with an ESR of 20 mΩ and an ESL of 0.1 nH, respectively. The parallel arrangement gives an effective ESR and ESL of 1 mΩ and 5 pH, respectively, and a total capacitance of 16.4 μF.

Figure 9 shows the impedance curve, labeled '+2nd stage', computed with the power supply and both stages of decoupling inserted into the SPICE model. The portion of the circuit to the left of point *x* in Figure 8 was simulated. The upper cut-off frequency using both decoupling stages is approximately 200 MHz. The curve also contains two small anti-resonant peaks at approximately 6 KHz and 2.5 MHz. The choice to use a total capacitance value greater than 2 times the required (as estimated by Equation 7), helps with keeping the anti-resonance peaks below the target impedance of 6.25 mΩ.

The elements of the model shown between the labels *x* and *y* in Figure 8 are designed to model the connection from the power plane in the PCB down through the C4 connection to the CUT. Since there are usually multiple C4 power supply connections (not shown), it is not necessary that this network meet the target impedance of 6.25 mΩ by itself. (A closer inspection reveals that this is not possible given the 250 mΩ contact resistance shown in Figure 8.) However, the equivalent impedance given by the parallel connection of all C4s must meet the target impedance. This sets a lower bound on the number of V_{DD} C4 ports required.

The V_{DD} C4 network model contains 6 elements. The left-most R and L identified as 'Probe card' in Figure 8 represent the resistance and inductance associated with the *via* used to route power from the power plane in the PCB to the membrane probe. Although the resistance of the *via* is small at 2 mΩ, the series inductance is significant. For standard 63 mil PCB cards, the *via* inductance can be calculated using Equation 8 [14]. Here, L_{via}, h and d are the

$$L_{via} = 5.08h\left[\ln\left(\frac{4h}{d}\right) + 1\right] \quad (8)$$

inductance, height and diameter of the *via*. In our model, we assume the power plane is positioned on the layer above the ground plane which occupies the lowest inner layer of the PCB (ground plane is not shown in Figure 7). We have also assumed that the parasitics associated with the routing of PWR and GND down to these planes is negligible. The model for a 10 layer board described in [14] gives a typical *via* diameter value of 16 mil and a height of 5 mil for the GND plane and 11 mil for the PWR plane. Plugging these values into Equation 8 yields an inductance of 31 pH from the GND plane and 112 pH from the PWR plane. The model for the C4 membrane probe was obtained from [16] and [17]. The 250 mΩ C4 contact resistance is an average value for "lead free" C4s.

The curve labeled '+C4 model' in Figure 9 gives frequency response of the PDS when 64 power C4s are included in the model. (Note that we did not include ground C4 elements in the simulation model. The complete model with GND C4s is analyzed in the next section). The parallel arrangement of 64 C4s yields an effective resistance of approximately 3.9 mΩ and adds to the 1 mΩ power supply series resistance. This 4.9 mΩ equivalent resistance defines the impedance amplitude at lower frequencies as shown in the left-most portion of the +C4 model curve in Figure 9. Interestingly, the anti-resonance peaks introduced by the probe card decoupling model cause the *+C4 model* curve to extend above the target impedance in frequency bands 3-20 KHz and 1-4 MHz.

The *+C4 model* curve exceeds the target impedance beyond 100 MHz. The on-chip decoupling capacitance needs to take over at this point. The curve labeled '+On-chip decouple' was generated from a simulation in which a lumped decoupling capacitance was added to the model. This capacitor is shown in the center portion of Figure 8 along with the current source used to stimulate the circuit. Except for the appearance of a third anti-resonance peak in the 70-300 MHz frequency band, the on-chip decoupling capacitance keeps the total impedance below the target for higher frequencies.

The use of a lumped model for on-chip decoupling capacitance is only appropriate for a first-order analysis. The resistance characteristics of the chip's power grid reduce the effectiveness of on-chip decoupling. The model

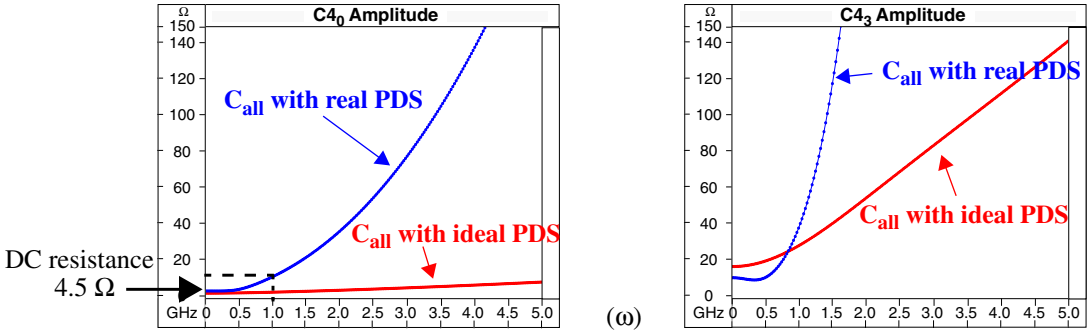

Figure 10. Impedance amplitude characteristics at $C4_0$ and $C4_3$ with the PDS model computed using an AC sweep of a current source tied to M1 at intersection of v-slice and h-slice lines in Figure 1(a).

analyzed in the next section replaces the lumped capacitance with the RC model of the Quad.

6.0 Impedance profile of the power grid with PDS

In order to determine the impact of the PDS on the measurement resolution of the CUT's power supply transient signals, an AC simulation was performed on the combined PDS RLC model and the power grid C_{all} RC model described in Section 4.0. The PDS model includes all the elements to the left of point *y* in Figure 8. The far right side of Figure 8 shows a block that represents the RC model of the Quad. The Quad's RC model replaces the lumped capacitance shown between points *y* and *z* in the Figure 8. The Quad includes 4 V_{DD} ports and 6 GND ports. Therefore, the *Probe card* and V_{DD} and GND *C4 model* elements shown in Figure 8 are replicated in the simulation model to accommodate these grid connections.

One of the objectives of this research is to determine the frequency range that is measurable using external test equipment. It follows from the analysis presented in the previous section that measurement points close to the chip are likely to enable the measurement of a wider range of frequencies. The closest measurement point possible is identified at point *A* in Figure 8. In the physical model shown in Figure 7, this point is located in the lowest layer of the PCB, very close to the actual membrane probe.

The results of the simulations are shown in Figure 10 using the same scaling factor as those shown in Figure 5. For ease of comparison, the C_{all} amplitude curves from Figure 5 are repeated in Figure 10 and are labeled 'C_{all} with ideal PDS'. The amplitude curves under the combined model are labeled 'C_{all} with real PDS'.

The curves in both of the amplitude plots show that the *real PDS* attenuates the high-frequency components of the transient signals starting at approximately 500 MHz. For example, the DC attenuation value is nearly tripled at 1 GHz for $C4_0$, as shown by the dotted line in the left plot of Figure 10. The passive elements in the PDS in combination with the on-chip decoupling capacitance are responsible for this attenuation. Therefore, frequencies above 1 GHz are measurable but will be attenuated by a factor greater than 3 over the grid's resistance attenuation factor. It follows from this analysis that any type of defect detection or localization technique must necessarily make use the lower frequency components in order to be practical.

The phase spectra from this analysis are not shown since they are similar in shape to the C_{all} phase spectra shown in Figure 5. The most notable difference is the increase in the maximum value of phase to approximately 200 degrees.

6.1 Time domain analysis

The impedance profiles described in the previous section illustrate the effects of the PDS and power grid RLC elements from the frequency domain perspective. This section focuses on analyzing the effects of these modeling elements on the time domain waveforms.

Figure 11 shows two sets of time domain I_{VDDx} waveforms, both generated under the C_{all} capacitance model described in Section 4.0. The PDS was excluded from the simulation model for the waveforms shown on the left, labeled 'Ideal PDS', while it was included for those shown on the right ('Real PDS'). The waveforms are positioned in each plot in a manner consistent with the position of the C4 from which they were measured (see Figure 1). The transient current stimulus is a piece-wise linear triangle waveform 1 ns wide and 20 mA high connected between two M1 PWR and GND runners at the position given by the intersection of the v-slice and h-slice lines in Figure 1.

The impedance characteristics of the PDS *reduce* the magnitude of the spatial variations in the C4 currents over that provided by the grid alone. The areas under each of the waveforms are shown in the plots of Figure 11 to illustrate this effect. For example, the areas under the waveforms from the *Ideal PDS* simulation model for $C4_0$ and $C4_3$ are 5.3 and 1.0, respectively. In contrast, the waveform areas computed under *Real PDS* model are 4.1 and 1.6, respectively. The ratio of these areas under the two models for each pairing of the C4s reflects the reduction in the spatial variations. For example, the ratio $C4_0/C4_3$ for the *Ideal PDS* areas is 5.3 while the ratio for the *Real PDS* areas is 2.6. The C4 contact resistance is the primary factor for the decrease in regional signal behavior. This is reflected in the

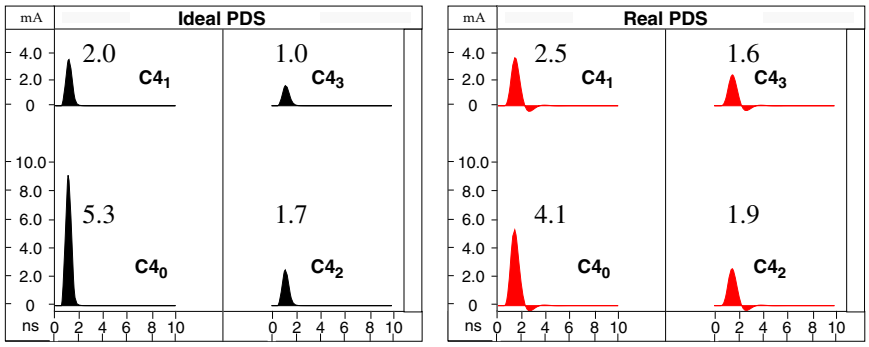

Figure 11. Time domain analysis using ideal and real PDS model.

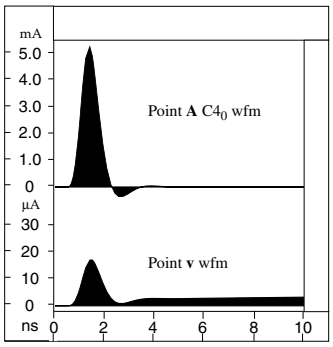

Figure 12. Single vs. multiple C4 measurement points.

frequency domain analysis as well. For example, in the amplitude plot for $C4_3$ shown on the right of Figure 10, the DC resistance value for the *Real PDS* curve is lower than the DC value for the *Ideal PDS* curve.

The I_{DDT} testing methods identified in Section 2.0 are based on the analysis of a single I_{DDT} value per vector. The most likely point of measurement is given by point *v* in Figure 8. The top waveform in Figure 12 is the time domain I_{DDT} waveform (shaded to a zero baseline) generated at our proposed measurement point *A* while the bottom waveform is generated at point *v*. The peak magnitude of the *point v* waveform is over 3 orders of magnitude smaller and contains no regional information. This analysis suggests that accessing the transients at measurement points between the power plane on the PCB and the C4 connections to the chip enhances the resolution of the core-generated transients and provides regional signal information not available in single point measurement techniques.

7.0 Conclusions

This research investigates the characteristics of the RC model of the supply grid and the RLC model of the PDS as a means of validating the practical aspects of transient-based test and fault localization techniques. The impedance characteristics of the combined system indicate that the combination of the PDS RLC model and power grid decoupling capacitance makes it possible to measure frequency components of the power supply transient signals up through 1 GHz without significant attenuation if the measurement points are close to the CUT. Although the upper bound on the measurable frequency range is limited, previous work has demonstrated that this level of frequency resolution is sufficient for fault detection and localization using power supply transient testing methods.

Acknowledgments

We thank Anne Gattiker, Sani Nassif and Dennis Conti at IBM for their support of this research.

References

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

[2] A. Germida, Z. Yan, J. F. Plusquellic and F. Muradali, “Defect Detection using Power Supply Transient Signal Analysis”, *ITC*, pp. 67-76, 1999.

[3] J. F. Plusquellic, D. M. Chiarulli, and S. P. Levitan. “Identification of Defective CMOS Devices using Correlation and Regression Analysis of Frequency Domain Transient Signal Data,” *ITC*, pp. 40-49, 1997.

[4] J. F. Frenzel and P. N. Marinos, “Power Supply Current Signature (PSCS) Analysis: A New Approach to System Testing”, *ITC*, pp. 125–135, 1987.

[5] S. Su and R. Makki, "Testing Random Access Memory by Monitoring Dynamic Power Supply Current”, *JETTA*, Vol. 3, No 4, pp. 265-278, 1992.

[6] J. S. Beasley, H. Ramamurthy, J. Ramirez-Angulo, and M. DeYong, “IDD Pulse Response Testing of Analog and Digital CMOS Circuits”, *ITC*, pp. 626–634, 1993.

[7] M. Sachdev, P. Janssen, and V. Zieren, “Defect Detection with Transient Current Testing and its Potential for Deep-Submicron CMOS ICs”, *ITC*, pp. 204-213, 1998.

[8] B. Vinnakota, W. Jiang and D. Sun, “Process-Tolerant Test with Energy Consumption Ratio”, *ITC*, pp. 1027-1036, 1998.

[9] K. Muhammad and K. Roy, “Fault Detection and Location using Idd Waveform Analysis”, *IEEE Design and Test of Computers*, Volume 18, Number 1, pp. 42-49, 2001.

[10] I. de Paul, M. Rosales, B. Alorda, J. Segura, C. Hawkins and J. Soden, “Defect Oriented Fault Diagnosis for Semiconductor Memories Using Charge Analysis, Theory and Experiments”, VTS, pp. 286-291, 2001.

[11] MOSIS TSMC’s 0.25um process information at http://www.mosis.edu/Technical/Testdata/tsmc-025-prm.html

[12] C. Patel, E. Staroswiecki, S. Pawar, D. Acharyya, and J. Plusquellic, ”Diagnosis using Quiescent Signal Analysis on a Commercial Power Grid”, *ISTFA*, pp. 713-722, 2002.

[13] A. Singh, J. Plusquellic and A. Gattiker, “Power Supply Transient Signal Analysis Under Real Process and Test Hardware Models”, *VTS*, pp. 357-362, 2002.

[14] H. Johnson and M. Graham, *High-Speed Digital Design: A Handbook of Black Magic*, Prentice Hall, Upper Saddle River, New Jersey, 1993.

[15] L. Smith, R. Anderson, D. Forehand, T. Pelc, T. Roy, “Power Distribution System Design Methodology and Capacitor Selection for Modern CMOS Technology”, *Trans. on Advanced Packaging*, Vol. 22, Number 03, pp. 284, Aug. 1999.

[16] A. Barber, K. Lee, and H. ObermaierBarber, “A Bare-chip Probe for High I/O, High Speed Testing”, Hewlett Packard *HPL 94-18*, March, 1994.

[17] S. McKnight, “Probing Lead Free Solder Bumps in Final Wafer Test”, presentation at Southwest Test Workshop, June, 2002.