Critical Timing Analysis in Microprocessors Using Near-IR Laser Assisted Device Alteration (LADA)

Jeremy A. Rowlette and Travis M. Eiles
Intel Corporation

Abstract

A scalable laser-based timing analysis technique we call laser assisted device alteration (LADA) is introduced for the rapid isolation and analysis of defect-free performance limiting circuits in advanced flip-chip packaged microprocessors and other complex IC's. The technique, which has been demonstrated to be widely applicable to production level as well as motherboard/system level testing, uses a laser incident from the backside to perturb the timing of internal nodes by means of temporary alteration of transistor characteristics primarily by means of localized photocurrent injection. The relevant physics describing the effects of near-IR laser sources on modern day CMOS FET devices and circuits is discussed in this paper in the context of achieving precision picosecond-scale timing adjustment. A selected case study where this technique was used to isolate a critical path circuit in a leading edge 130 nm generation product is provided. Scaling trends for LADA and other relevant issues are discussed.

I. Introduction

As microprocessor core operating frequencies continue to accelerate into the microwave frequency range, challenges with on-die synchronization increase accordingly. To the extent that on-die synchronization limits the upper frequency bound of a complex IC, a significant portion of the silicon validation cycle is devoted to isolating and fixing the performance-limiting circuits [1].

Precision through-substrate, or backside, probing technologies such as optical probing [2, 3] and time resolved emission among others* [4, 5] in conjunction with focused ion beam (FIB) circuit edit technology [6] have been developed for the purpose of analyzing and verifying the root cause of performance-limiting signal paths. However, these analytical technologies are fundamentally limited by their serial measurement execution. That is, the group of logic gates comprising a suspected critical signal path must be hand selected prior to any measurement being made. Despite increasing sophistication of design-for-test (DFT) features such as *scan* [7, 8] and on-die clock shrinking [9, 10], which help narrow the list of failing node candidates, there is still a need for the ability to isolate marginal node(s) to a small group of logic gates in an efficient and highly reliable manner. Soft Defect Localization (SDL) [11], an expansion of the Resistive Interconnect Localization (RIL) technique [12] has shown good results for isolating non-catastrophic defects or "soft" defects. Although we find the SDL methodology to be useful in semiconductor failure analysis, we find that it does not yield compelling results in the debug of circuit design marginalities, a gap which this paper intends to address.

This paper introduces a laser-based analytical technique, referred to as laser assisted device alteration (LADA), which is qualitatively similar in form to SDL but which has fundamental differences in both approach and application. Unlike SDL, LADA provides the ability to rapidly isolate critical signal paths or "speedpaths" and their limiting components, which are assumed to be absent of manufacturing defects†, down to the individual logic gate level with high confidence while scanning large (>2mm^2) regions of the die. The differences between the LADA and SDL techniques will become more evident in the following sections.

Once the critical node(s) have been identified using this technique, further analysis is typically required to determine the precise relationship of the node(s) to the critical timing event whereby analytical debug technologies previously mentioned are well suited.

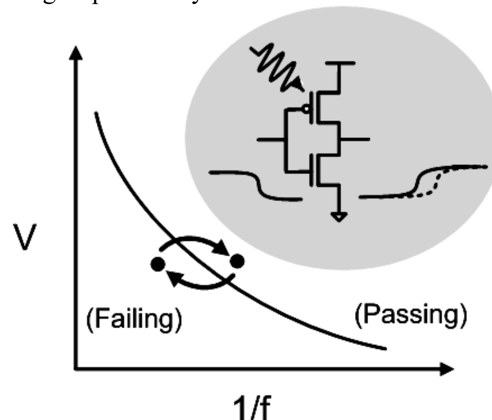

Figure 1. The LADA concept: A scanning laser perturbs critical path transistor characteristics thereby changing signal timing sufficiently to cause transitions in the DUT's pass/fail behavior. The result is the precise x-y location of performance limiting gates determined in a matter of minutes.

* For example: e-beam probing, high-speed contact probing, and infrared emission microscopy (IREM).

† There are no electrically observable defects in the critical signal path.

 0-7803-8106-8/03 $17.00 Copyright 2003 IEEE

II. LADA

Despite differences in the details of the implementation and the underlying physics, the concept of the technique presented in this paper was first explored in practice by Pronobis and Burns [13,14] in the early 1980's and later by others [15] using pulsed lasers operating in the visible range. The basic concept is as follows. With the DUT biased at a marginal state, i.e., close to the pass/fail boundary for a fixed temperature, voltage, and frequency setting, a laser perturbs the timing of a node located within the critical signal path such that the maximum operating frequency of the DUT changes by an observable amount. With existing standard IC test platforms, this change in upper frequency bound is easily observable. One elegant methodology for observing this frequency shift "on the fly" was presented in the initial demonstration of RIL by Cole *et al.* [11], which we also employ in the implementation of LADA. The one addition we make is the use of an asynchronous "jamb" latch in the trigger path, which extends the ability to operate the scanning clock asynchronously with respect to the DUT clock*. Figure 1 depicts the general concept just described. A typical voltage-frequency *shmoo* plot is shown in which there is a distinct boundary between the passing and failing state of the DUT. When operated in close proximity to this boundary, only small perturbations in the critical signal path timing are necessary to cause a transition in the pass/fail state. Depending on where the DUT is biased and the type of timing perturbations induced, e.g., increase or decrease in delay, the transition can either be *pass-to-fail* or *fail-to-pass*. In typical analyses, both trigger polarities along with the two corresponding bias settings are used for each case making no assumptions of the type of edge shifting to be induced in the silicon.

In LADA, a continuous wave (CW) near-IR wavelength laser is either scanned continuously across the region of interest as the tester exercises the DUT repeatedly with a single test pattern or is precisely positioned on a suspected node of interest in what we refer to as the "parked" mode. The latter technique is particularly suited for testing out hypotheses quickly and is effectively the equivalent of a "soft FIB edit".

The tester platform is not limited to production level IC testers. Motherboard/System level testing is an alternative, low-cost implementation scheme, which has been demonstrated, albeit with some caveats, which we will not discuss in this paper†. During each test cycle, the pass/fail state of the DUT is interrogated and a binary output trigger indicating the pass/fail state of the DUT is generated by the tester. This pass/fail trigger is then sampled by the data acquisition electronics of the LADA system, which internally synchronizes the trigger event to the laser spot position on the die. Correlations between changes in the pass/fail state of the DUT and the laser spot are determined as "sensitive" nodes that are bucketed for further investigation using more refined LADA techniques such as precise positioning of the beam on a single node in addition to complementary probing techniques previously listed. Figure 2 shows a high-level block diagram of the production system used to implement this technique, which was constructed in partnership with Checkpoint Technologies, LLC (San Jose, CA).

The amount of timing perturbation at a given node is directly related to the laser power. Thus, by fine-tuning the laser power incident to the DUT, one can tailor the amount of time shifts incurred in the silicon and therefore determine relative sensitivity or marginality among a large number of nodes in a highly parallel fashion. With continued increases in operating frequencies, synchronization issues are measured on an ever-diminishing timescale presently marked by tens of picoseconds or less. Time-shifts on this scale are well within the range achievable using LADA as will be shown. It will also be shown that decreasing the laser power down to a minimum threshold level distinguishes the most sensitive node within the scan field, where it is highly probable that this node is the performance-limiting element for the given test pattern. In addition to isolating the most sensitive node in the signal path, large portions, if not all, of the critical signal path can be mapped out with a single scan of the laser by increasing the laser power above this minimum power threshold level. We will return to some of these statements in our discussion but first a more detailed description of the methodology and the underlying physics of the interaction of the laser with modern day CMOS devices and circuits is necessary.

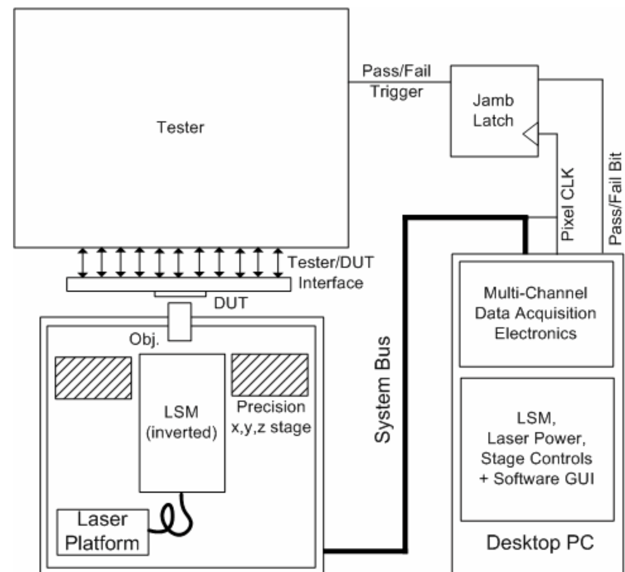

Figure 2. Block diagram of the LADA system. The LADA system is fully integrated into a production diagnostic tool, which docks directly to standard IC test platforms.

* The jamb latch ensures that all data is captured by the LADA system.
† Exceptionally long loop lengths (>10s) are often necessary in system level testing thus making it practically difficult to run continuous laser scans in search of the critical path device(s) as it is required that 1 pixel effective dwell time be ≥ length of the test loop.

III. Theory of LADA

The basic operating principle of LADA is that a laser temporarily* perturbs transistor characteristics such that the propagation time of signals traversing the illuminated device(s) changes in a controlled and repeatable manner. There are two basic modes of operation available for inducing such non-permanent alterations in device behavior: (1) using the laser as a localized heat source, or (2) using the laser as a localized current source. Either of these two modes is enabled through proper selection of the laser wavelength. CW Nd:YVO_4 and Nd:YAG, with wavelengths tuned to 1340 nm and 1064 nm respectively, are laser sources that are commercially available and routinely used in debug, fault isolation and failure analysis labs throughout the industry. With its longer nominal wavelength, Nd:YVO_4 is well-suited for localized heating whereas Nd:YAG is a good source for photocurrent in silicon devices, while also enabling a slightly improved spatial resolution. Both lasers are in the near-IR band and have sufficiently low absorption coefficients for imaging through at least 100 μm of heavily doped p+ silicon†. In all cases the same laser that is used to induce spatially localized timing perturbations is also used for optical imaging of the transistor layer for navigation and localization.

Although localized heating of "healthy" transistors, i.e., transistors free of process-related defects or marginalities, will induce a corresponding time shift in the DUT, the time shifts are typically very weak (< 2ps). Therefore, despite statements made in the literature regarding undesirable photocurrent effects [12], we find that the photocurrent mode of operation is the preferable approach for generating practical timing-shifts in today's technologies for circuit debug applications. This is one of the main distinguishing features between the SDL and LADA approaches.

A. Perturbation of CMOS Device Characteristics

Localized photocurrents are generated within the active regions of the transistor layer using a Nd:YAG CW diode pumped laser which has a nominal operating wavelength of 1064 nm. This wavelength corresponds to a photon energy of ~1.17 eV, which *is* sufficient to cause electron-hole (E-H) pair generation at typical DUT operating temperatures. The photo-generated currents induced through local carrier generation affect the timing at a node either directly by assisting in the charging or discharging of a load capacitance or indirectly through the local modulation of electric potential thus changing the DC operating point of the transistor.

The effects of the 1064 nm wavelength laser were studied extensively for short-channel n and p-channel transistors fabricated in twin well CMOS process technologies. Complete sets of I_D-V_{DS} and I_D-V_{GS} curve sweeps were made under a range of laser illumination conditions. Referring to Figure 3, drift currents may be generated within any and all of the exposed reverse-biased junctions including the transistor's source/drain junctions and the n-well/p-substrate junction enclosing the PMOS device. However, of particular importance are the photocurrents generated in and collected by the drain junction of the NMOS device as well as the n-well/p-substrate junction of the PMOS device, which are explicitly shown in Figure 3. Recall that the direction of illumination of the incident laser is from the backside of the die. Also note that the substrate has been mechanically thinned to approximately 100 μm and a thin film anti-reflection coating has been deposited.

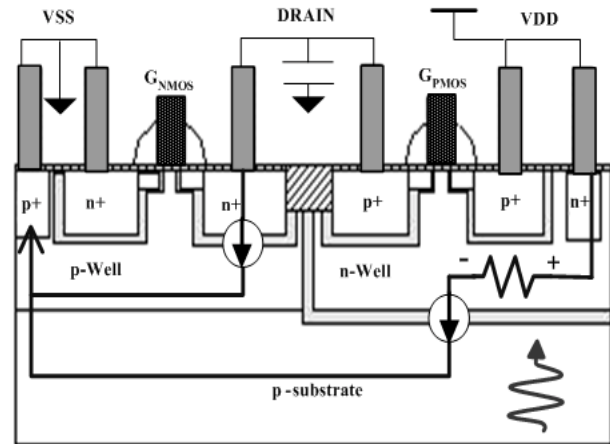

Figure 3. Critical photocurrent sources in CMOS with illumination from the backside.

Despite having the same phenomenological origin, experimental observation shows a distinct contrast between NMOS and PMOS device alteration in the presence of the 1064 nm laser beam when illuminated from the backside. As will become apparent, the exposure of the n-well junction plays a key role in causing this discrepancy. The NMOS drain current, I_D, exhibits a constant offset current proportional to the laser power, which is practically independent of the transistor biasing state. The additional drain current due to the laser flows from the drain node to the substrate contact through the substrate as depicted in Figure 3. Figure 4 shows the relationship between drain current and incident laser power, taken at $V_{GS} = V_{DS} = 0$ V bias, indicating an unmistakably linear relationship. This characteristic is identical to that of a photodiode, which is expected to have a linear response over the power range used. A variable constant DC current source in parallel with the NMOS device with a polarity as shown in Figure 3 and 4 easily models this effect.

The additional drain current is caused by the separation of electron-hole (E-H) pairs generated within the reverse biased drain junction as well as the collection of electrons generated in the p-type region (assuming low-level injection in the p+ background) which diffuse outward from the focusing cone of the laser. For the typical case that the laser spot is greater than the junction

---

* This statement is worth reiterating since this is a fundamental requirement for the technique. For typical laser powers required, no hysteresis in device characteristics was observed.

† The sample is prepared in a manner consistent with that described in reference 3.

area, the magnitude of the photocurrent for NMOS devices will be dependent upon the incident laser power and the area of the drain junction. It is important to note that the photocurrent source is always on and sinking charge off of the drain node while the laser is illuminating the device. This fact has important consequences in domino logic topologies whereby photocurrent magnitudes comparable to the current drive of the “keeper” transistors (on the order of 200 µA today) may cause unintentional switching in the perturbed logic path. Despite this potential drawback, we do not find it necessary to be concerned since it rarely limits the ability to use this technique. This is because the photocurrent injected into the keeper transistor is limited by its own drain junction area (assuming an NMOS device). The near minimum size of typical keeper transistors helps to minimize this undesirable effect.

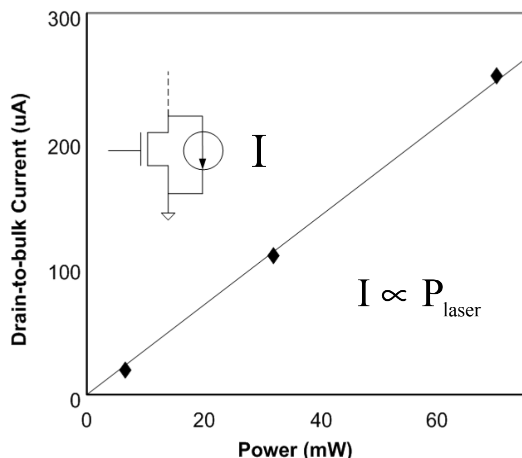

Figure 4. Additional drain current as a function of laser power measured at the objective lens for a 10 µm NMOS device fabricated on 0.25um technology with Vgs=0V.

Unlike NMOS devices, PMOS devices tend to exhibit negligible drain current offset. Therefore for all practical considerations, the current of the PMOS device contacts the horizontal axis (V_{DD}-V_{SD}) when the source-to-drain voltage is zero[*]. This behavior is seen in Figure 5 where the effect of increasing laser power is shown for a typical I_D-V_{DS} family of curves. Here, data is taken at 2 values of laser power (P_2 and P_1 where $P_2 > P_1$) in addition to the zero-illumination or nominal condition (P_0).

The discrepancy between NMOS and PMOS device alteration is due to the strong influence on device behavior associated with the photocurrent generated in and collected by the n-well/substrate junction. For the PMOS device, the dominant photocurrent source occurs in the n-well/p-substrate junction rather than in the source/drain junctions. Consequently, a significant photocurrent flows from the n-well tap to the p-substrate tap through the n-well and the n-well/substrate junction, which induces a potential drop between the n-well contact and the laser spot as depicted in Figure 3. If the laser spot is centered about the active region of the device, this potential drop increases the PMOS source-to-body voltage, V_{SB}, thereby reducing the threshold voltage magnitude, $|V_{t,p}|$, as a result of the well-known body effect [16]. This voltage drop increases with increasing laser power as a result of the proportional increase in photocurrent. A reduction in $|V_{t,p}|$ leads to an increase in the gate overdrive voltage and therefore increases the device transconductance, g_m, for a given source-gate voltage, V_{SG}.

The reduction in threshold voltage magnitude is more pronounced in Figure 6 where the drain current, I_D, is plotted with respect to the source-to-gate voltage, V_{SG}, for a fixed drain-to-source voltage magnitude, $|V_{DS}|$, of 1.5 V. The family of curves shown in Figure 6 was generated by incrementing the laser power in discrete steps and sweeping the device terminal voltages over a fixed parameter space. The family of curves shown in Figure 7 was obtained in similar fashion but in the absence of illumination and by manually decrementing the n-well voltage in steps of 100 mV using an independent voltage source. Figure 7 is provided purely for the purpose of comparing the effects of device illumination with the effects of back-gate biasing for PMOS devices. In each case, the magnitude of $V_{t,p}$ decreases with increasing laser power or similarly decreasing n-well voltage with respect to a fixed source potential.

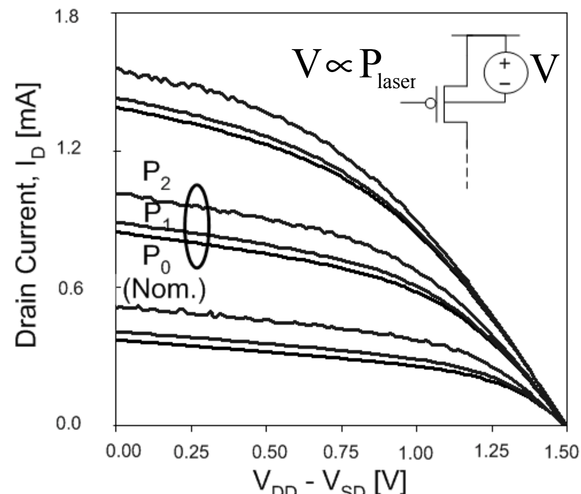

Figure 5. PMOS I_D-V_{DS} curve family for various source-to-gate voltages and laser illumination. $P_0 \rightarrow$ 0 mW; $P_2 > P_1 >$ 0 mW

Figure 8 shows the explicit relationship of the change in $V_{t,p}$ for both cases. Figure 8, in combination with Figures 6 and 7, shows that increasing the laser power has the nearly identical effect on PMOS devices as reducing the n-well voltage, to within a multiplicative constant. This is expected since the photocurrent generated in the n-well/substrate junction is linearly proportional to the laser power and the voltage drop is linearly proportional to the photocurrent[†].

[*] A fraction of the light is absorbed in the drain junction of the PMOS device, which contributes to the drain current. However, we find this contribution to be considerably less than for NMOS devices.

[†] The proportionality constant is simply the effective n-well resistance.

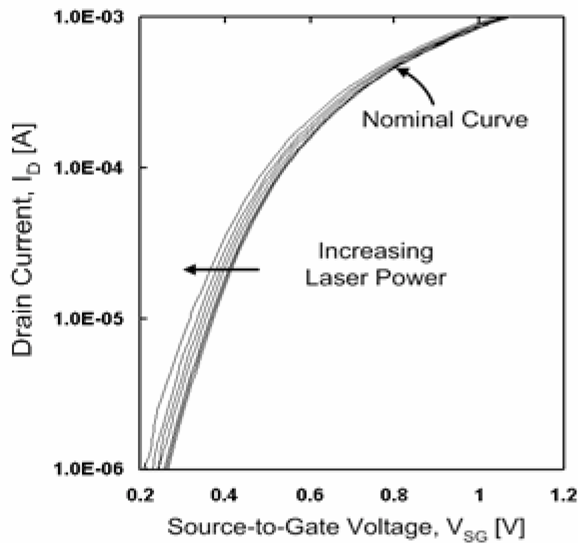

Figure 6. Perturbation of I_D-V_{SG} curves by 1064 nm laser for a p-MOSFET of 10 μm gate width fabricated on 0.25 μm technology.

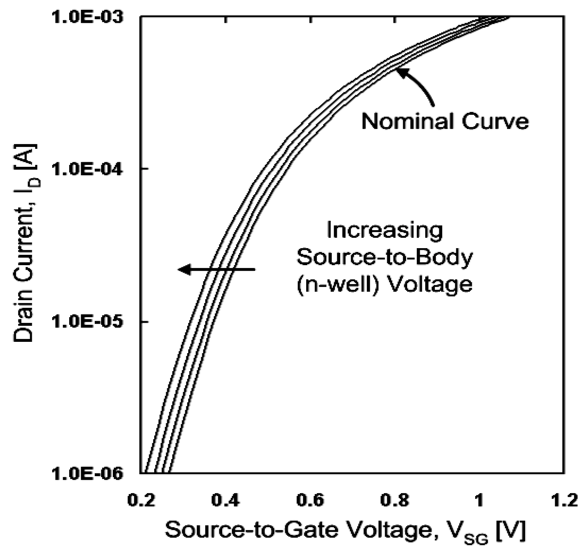

Figure 7. The effect of independent n-well biasing on I_D-V_{SG} curves for a p-MOSFET of 10 μm gate width fabricated on 0.25 μm technology provided for comparison purposes.

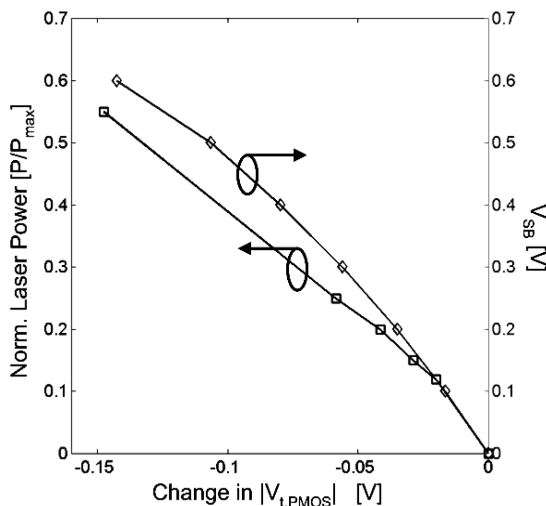

Figure 8. PMOS $|V_t|$ reduction due to reduction in n-well voltage, V_B, with respect to the PMOS source voltage, V_S (diamond) and due to increasing laser power (square) .

Therefore, this effect can be modeled by a variable DC voltage source inserted between the body contact and body supply voltage (typically maintained at V_{DD}) with a polarity shown in Figure 5. The effective n-well resistance will depend on the proximity of the device to the nearest n-well contact(s) and the relative position of the laser spot with respect to the device and the nearest contact(s). However, layout design rules for modern IC's typically mandate that n-well contact densities adhere to strict uniformity guidelines, which helps to minimize this variance greatly. In practice, this variability is rarely observed.

Localized heating of the transistor can also be used as a means for altering the timing of a node and can be achieved using a Nd:YVO_4 CW diode pumped laser operating at a center wavelength of approximately 1340 nm, also available in the production tool. Primarily, transistor heating reduces the local carrier mobility in the channel and source/drain region thus limiting the effective drive strength of the transistor. This causes a "slow down" of logic transition edges regardless of the edge polarity. However, because the heating is restricted to a localized region surrounding the focused laser spot and because the maximum temperature rise of the laser is restricted to a few tens of degrees at best [17], the magnitude of time-shifts is typically very weak compared to the nominal propagation delay of the node. Despite this weak effect, reports of limited use of this effect have been reported [11, 18]. Because the thermally induced timing effects are typically too weak to observe in "healthy" silicon, we almost exclusively operate in the photocurrent injection mode where appreciable time shifts can be induced. In techniques such as Soft Defect Localization (SDL), localized heating can result in significant timing shifts since the resistive defect dominates the interconnect parasitics and the temperature coefficient of the defect may be very large. In applications using LADA, the DUT is assumed to be free of such defects.

B. Precision Timing Alteration

Regardless of the underlying mechanism(s), the change in signal timing for a typical CMOS stage (static or dynamic) is determined by the net change in node current deliverable by the source

$$\Delta t_d = \frac{C_{node} V_{DD}}{\Delta I_{source}} \quad [s] \tag{1}$$

where C_{node} and V_{DD} are the effective node capacitance and supply voltage respectively. To first order, the effective node capacitance and supply voltage is unperturbed by the laser and thus we concern ourselves solely with the change in node current induced by the laser. We will focus on the photocurrent injection mode of operation because of its stronger significance to the LADA technique. However, before doing so, we briefly illustrate our motivation for placing less emphasis on localized heating as a method for altering nodal timing in

debug applications. A plot of the 50/50 propagation delay as a function of junction temperature for a typical inverter fabricated on 130 nm generation technology is shown in Figure 9. The dependency is approximately linear over typical chip operating temperatures and has a temperature coefficient of about 10^{-2} ps/°C.* This data is consistent with reports of very weak thermally induced shifts in ring oscillator frequency by reference 11. The relatively weak thermal effect is the primary reason for using the photocurrent injection mode of operation for achieving relevant magnitudes of spatially localized timing shifts.

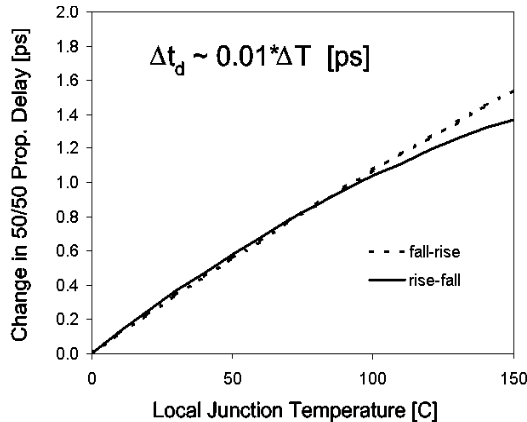

Figure 9. Simulation results showing change in 50/50 propagation delay for a typical 130 nm technology inverter vs. junction temp.

We now turn our attention to the more desirable photocurrent injection mode of operation and consider the effects that NMOS and PMOS device alteration have on the change in signal propagation delay. The additional current source in parallel with the NMOS device naturally strengthens the NMOS device. The additional photocurrent subtracts from the PMOS source current during low-to-high transitions and thus effectively weakens the PMOS device. Neglecting other considerations, this causes an increase in charging time of the output load capacitance when the logic signal is transitioning from low to high and a reduction in discharge time for high to low transitions.

In reference to the PMOS device alteration, reduction in the threshold voltage magnitude causes the PMOS transistor drive strength to be preferentially increased over that of the NMOS transistor. Therefore, neglecting all other considerations, low-to-high transition times decrease while causing the high-to-low transitions to increase. In addition, the modification of the threshold voltage adjusts the gate transition point, which changes the turn-on time of the illuminated pull-up devices. However, this has a minor effect on the overall edge transition time.

Timing structures located on a 130 nm generation test chip were used to quantify shifts in propagation delay with picosecond-scale resolution. The experimental setup for a series of precision timing alteration experiments performed is shown in Figure 10. For the experimental results shown in Figure 11, an isolated inverter located on a 130 nm generation test chip was used to study the effects of the 1064 nm beam on signal propagation delay. The results demonstrate that the change in propagation delay is approximately linear over the range of laser powers used in this experiment. For this experiment, the maximum laser power was about 400 mW, measured before the objective lens.

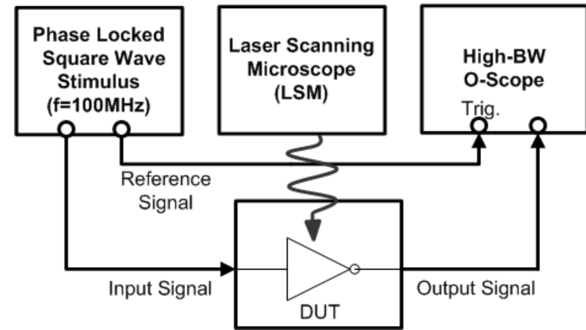

Figure 10. Experimental setup for measuring signal edge time shifts as a function of laser power and position.

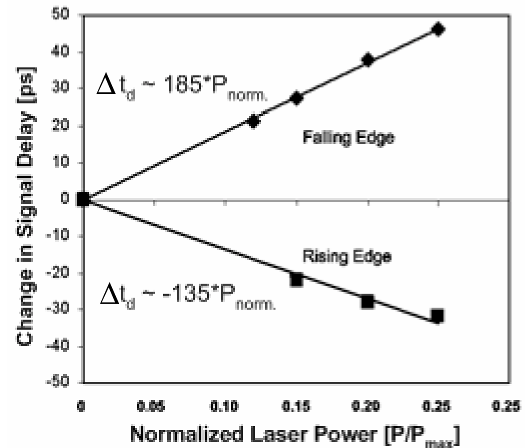

Figure 11. Time shift results from an experiment performed using the test setup shown in Figure 10.

Referring to Figure 11, the rising edge delay is reduced while the falling edge delay is increased. It is important to note that despite intentionally positioning the laser on the PMOS and NMOS devices exclusively, the result in signal propagation delays were unchanged. These results support two conclusions. The first conclusion is that for compact layout topologies, it is difficult to separate out the laser effects for both PMOS and NMOS devices. The second conclusion is that the PMOS device is the dominantly affected device in the CMOS pair provided that exclusive perturbation of either the NMOS and PMOS devices is not achievable. It is quite evident that the thermal diffusion of electrons generated in the bulk p-substrate increases the effective laser spot thus limiting the ability to isolate the LADA effect to a single NMOS or PMOS device for typical layout configurations. For a typical CMOS inverter most of the electrons generated in close proximity to the n-well junction of the PMOS

* The accumulation of propagation delays by logic gates that are exercised in series during a clock cycle gives rise to a larger, more significant, microprocessor frequency temperature dependency.

device will be collected by this junction and thus contribute to the n-well current. Because of the close proximity of the NMOS and PMOS device in typical CMOS layout topologies, it is difficult to separate out the two competing effects. The net result depends on the relative sizing of the two devices.

In the example shown, the relative sizing of the NMOS and PMOS devices was balanced to ensure near equal rise and fall times (equal drive current) and we found that the PMOS device dominates for this case. The dominance of the PMOS device over the NMOS device under equal illumination conditions is explained with a simple consideration of the how the MOSFET drive current is related to the threshold voltage. Consider a typical relationship between drain current and threshold voltage in the limit of velocity saturation:

$$(2) \quad I_D = v_{sat} C_{ox} W_p [V_{SG} - | V_{t,p} (V_{SB}) |] \quad [A]$$

where the threshold voltage, $V_{t,p}$, has been written as a function of the source-to-bulk voltage, V_{SB}. Based on the experimental results discussed in the previous section, this voltage is proportional to the product of the laser power and the effective n-well resistance (Eq. 3).

$$(3) \quad V_{SB} = I_{ph} \bullet R_{n-well} \propto P_{laser} \quad [V]$$

From Equations 2 and 3, it is apparent that the photocurrent, I_{ph}, generated in the n-well junction is actually amplified through the intrinsic transconductive gain of the MOSFET device. That is, small changes in the n-well voltage, produced by the n-well photocurrent, can lead to drain currents that are significantly larger than the photocurrent magnitudes themselves because of the normal voltage-to-current conversion process of the device. Also, because the n-well junction is typically much larger than the NMOS drain junction, more photocurrent is typically generated in or collected by the n-well/junction, which adds to the dominance of the PMOS device.

It is important to point out that the dominance of the PMOS device due to the laser illumination is really only valid under restricted conditions of laser power, circuit design parameters such as W_p/W_n ratio, and circuit topology, i.e., static vs. dynamic. An example of where the PMOS dominance does not always hold true presented briefly as an application note at the end of the paper.

IV. A Case Study on 130 nm Product

As demonstration of this technique, the results from LADA analysis performed on a prominent speedpath issue on a leading 130 nm generation microprocessor are shown. The critical schematic components of the speedpath are shown in Figure 12 and the layout of the functional block containing these components is shown in Figure 13. Above a threshold frequency and at a fixed ambient temperature and supply voltage setting, one of the data bit inputs to the six latches shown arrives late with respect to the falling edge of the clk# signal, leading to a system hang. This is a fairly generic scenario, on which many on-die synchronization issues can be modeled. The clk# signal, which controls the sampling of the 6 data channels is highlighted with a small, rectangular, white box in Figure 13.a and 13.b.

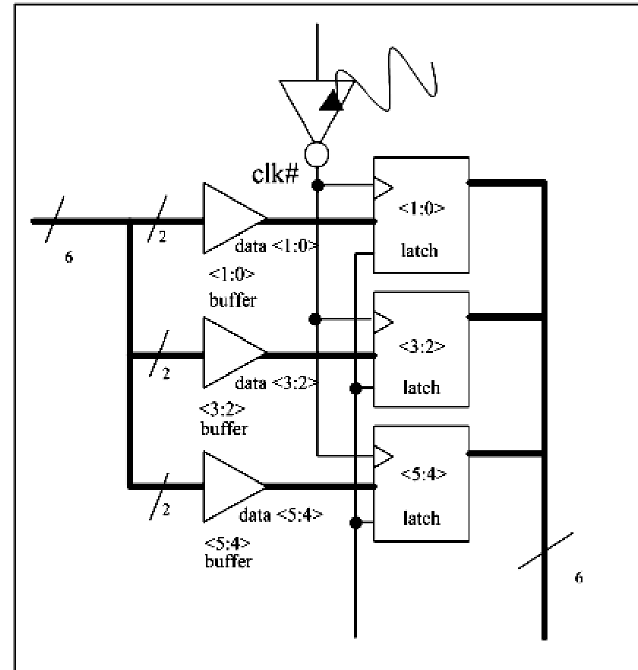

Figure 12. Schematic representation of the speedpath discussed in the 130 nm generation product case study.

An attempt to alter the timing of the clk# or critical (slow) data bit signals should cause the frequency performance of the chip to change in a consistent manner. For example, if the data path is made faster or the clk# signal made slower, then the upper frequency bound of the DUT should shift towards higher frequency. Likewise, if the data path were to be made slower or the clk# signal to be made faster, then the upper frequency bound should shift towards lower frequency. This behavior is clearly demonstrated using LADA. With the DUT biased in a nominally failing state near the pass/fail boundary, LADA analysis reveals a strong consistent signature from the inverter driving the clk# signal upon scanning a region approximately 2 mm^2 in area. The results are shown in Figure 14.b where the white highlighted pixels indicate the precise X-Y laser spot location where the tester registered a change in the pass/fail state of the DUT.

The entire field of view in Figure 14.a is approximately 0.026 mm^2. The highlighted pixels in Figure 14.b indicate that the DUT transitioned from a nominally failing state to a passing state when the laser was positioned at these pixel locations. The results are consistent with the data taken on the 130 nm generation test chip (Figure 11), where the falling edge was delayed. By sufficiently delaying the falling edge of the clk# signal, the slow data bit was able to arrive at the latch in time thus causing a shift in the frequency of the DUT towards higher frequencies.

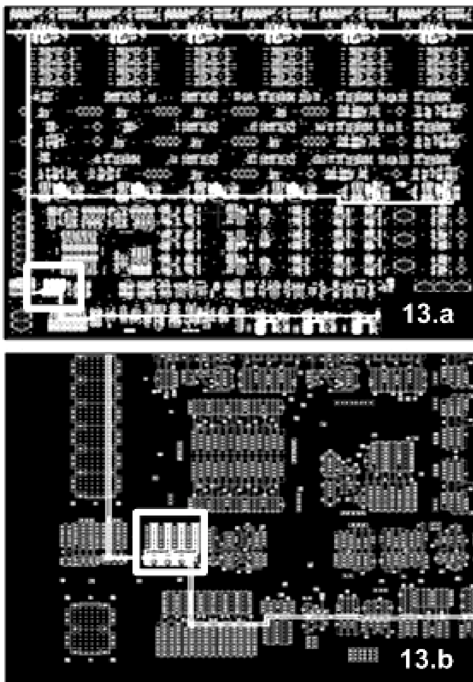

Figure 13. Layout of the region housing critical elements of the example speedpath. The rectangular region encloses the critical path inverter driving the CLK# signal.

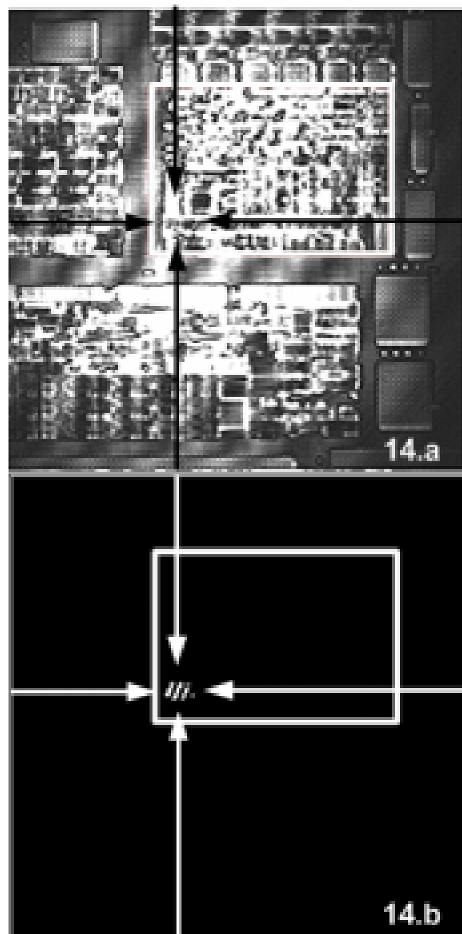

Figure 14. LADA data taken on a prominent 130 nm generation microprocessor. (14.a) The reflected image of the region of interest. (14.b) The Pass/Fail trigger data collected from the tester. The white highlighted pixels in 14.b indicate the precise position of the laser when the state of the DUT transitioned from a nominally failing state to a passing state.

V. Discussion

A. Practical implementation of LADA

Although several proposals and attempts have been made over the years to drive some form of this technique into mainstream IC diagnostics [13-15], we believe that it is only recently that such techniques have become viable. In particular, the development of near-IR laser probing technologies for backside (through-substrate) analysis have sidestepped the need for illuminating the transistor layer through a dense metal interconnect stack, which can consist of 7 or more metal layers [20]. Furthermore, logic edge transition times have shrunk by orders of magnitude since the early 1980's and are typically in the range of 10-50 ps today, made possible by a continued reduction in node capacitance and supply voltage. As a result, the instantaneous energy required for perturbing the timing of internal nodes by an appreciable amount has also diminished greatly enabling the use of continuous wave (CW) lasers. This ability to use CW lasers dramatically simplifies the test platform setup whereby the subtle issues of synchronizing mode-locked lasers to testers with sizable synch pulse jitter are entirely avoided [3]. In particular, the ability to operate the technique entirely asynchronously can be performed with the aid of the jamb latch, which ensures that the pass/fail trigger output from the tester will always be sampled regardless of the differences in frequency at which the LADA system (laser scan and data acquisition) and the tester operate.

B. Scaling of near-IR LADA

LADA is practically immune to the challenges associated with increasing operating frequencies. Because the laser is scanned at a relatively slow rate compared to the test loop, it acts as a DC source of perturbation and thus is indifferent to how fast the signals on the chip are operating. Therefore, LADA is expected to scale completely with respect to the frequency scaling requirements. We now turn our attention to the scaling of the LADA effect in relation to continuing dimensional process scaling.

Fortunately, the resolution of LADA is not as tightly coupled with the beam spot size as other probing technologies [2,3]. The fundamental diffraction limit does limit the image resolution and therefore degrades the ability to localize the sensitive node below this value, which is set at

$$R = \frac{0.61\lambda}{\text{N.A.}} \sim 0.76\,\mu\text{m} \quad (4)$$

where we have assumed the use of a high-N.A. objective lens optimized for through-silicon analysis*. This resolution can be enhanced through increases in N.A. with achievable values near and even beyond unity for

* We are assuming the use of a Nikon 0.85 N.A. lens with a 1 mm working distance.

liquid and solid immersion type lenses [20]. However, even if one were to reduce the spot size by an arbitrary amount, one would not generally increase the resolution of the LADA effect because the effect is limited primarily by thermal diffusion. Figure 15 illustrates this point. One way to decrease the effective spot area is to reduce the laser power. This reduces the local minority carrier concentration gradient and consequently the rate of carrier spreading out from the laser spot. This is demonstrated in the sequence of LADA images shown in Figure 16 where the laser power is incrementally reduced. As the laser power decreases, so does the *effective spot size* or rather *region-of-influence*. A high substrate doping concentration helps to restrict the carrier spreading to within a relatively short distance from the focused beam because of the reduced diffusion length.

Although reducing the laser power helps to increase the spatial resolution of the LADA effect, the reduction in laser power generally reduces the magnitude of the time shift as well, creating a window for optimization. To compensate for the diminished time shifts, the timing margin may have to be reduced accordingly to maintain test *observability*[*]. However, as the timing margin is reduced through manipulation of the local clock frequency, the stability of the bias condition tends to degrade causing the signal-to-noise ratio (SNR) to degrade. If the frequency of the noise is low compared to the scan rate then the SNR can be boosted through averaging over multiple laser scans over a fixed region of the die. Currently, we are not limited by the spatial resolution for most cases and so larger laser powers are acceptable despite the > 1 µm effective spot size. This is understandable if we consider the type of localization that is required. The desired outcome of LADA is to narrow the failing node(s) down to a single logic gate or small group of logic gates. Even with critical dimensions approaching or even exceeding 50 nm, the actual area consumed by most logic gates is still routinely on the order of 1 um or larger even in today's most advanced technologies. Furthermore, the need to distinguish between two adjacent logic gates is rare since logical and architectural insight can typically resolve such issues. At the very least, it should be considered a major feat to have narrowed down the critical node to between two gates on a chip comprised of tens or hundreds of millions of logic gates. At this point, precision timing measurement techniques such as optical probing (LVP) and time resolved emission (TRE) could take over in determining the precise nature of the critical signal(s).

C. Application Notes

One of the main challenges to practically implementing LADA into the mainstream debug environment is in understanding how to interpret the data because the analysis requires an understanding of the fundamental interaction of the laser with modern devices and circuits.

[*] This is a consequence of the binary behavior of this and the SDL technique. Either the time shift is greater than or less than the test margin for the given bias conditions.

During the course of developing and refining this technique subtleties have arisen, which are worth mentioning. We offer a few notes for consideration.

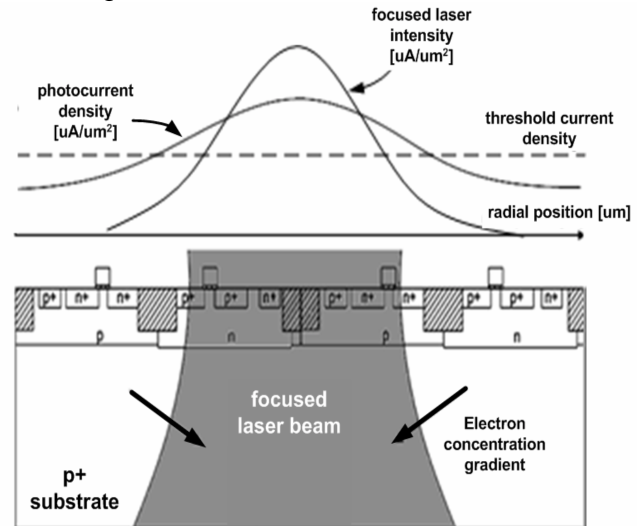

Figure 15. Depiction of the effects that diffusive currents have on the expansion of the effective beam spot or rather the effective influence area of the beam.

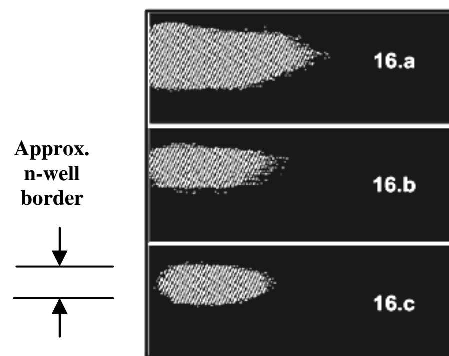

Figure 16. Reduction of laser power demonstrating corresponding reduction in effective beam spot. Laser power was highest for (a) and lowest for (c). The data was taken on the same critical path inverter discussed in the 130 nm generation case study.

In this paper, we have shown the effects of the perturbing laser on static logic gates. However, it is common for different types of logic topologies to be implemented in the same product. For example, dynamic logic circuits are found routinely in regions where speed is critical. For these types of circuits, independent NMOS and PMOS perturbations can be observed.

Another important consideration is that PMOS devices may not always dominate in the presence of the laser illumination, as we alluded to before. Though the data from the reported case study is consistent with data taken on numerous debug cases on a range of 130 nm generation as well as 90 nm generation products, there are cases that deviate from this trend. For example, if the gate is highly skewed to ensure a fast falling edge by making the pull-down NMOS network dominant, then the laser illumination may actually enhance the NMOS device more than the PMOS device at low laser powers. However, in most cases, as the laser power is increased, the PMOS device will tend to dominate because of the

non-linear dependence on threshold voltage shift in cases where the drain current is not entirely operating in the velocity saturation limit. This is demonstrated with the example shown in Figure 16 where the rising edge is actually delayed at lower power levels. At increasing laser powers, the rising edge delay reaches a maximum and then begins to decrease until there is eventually a net reduction in rise time delay. This gate was skewed such that the PMOS and NMOS device have equal gate width.

VI. Conclusion

Increasing chip complexity, measured as ever-increasing transistor density in combination with aggressive frequency scaling and the implementation of highly parallel architectures, inevitably drives engineering debug time to increase. At odds with the need for increased debug resources and validation time are intense market pressures to drive down the cost and duration of this cycle. This task can only be accomplished with advances in test technology that enable significant enhancements in capability and efficiency. We have demonstrated a powerful technique for efficiently isolating critical performance limiting nodes in complex integrated circuits, which are free from manufacturing defects. The LADA technique has proven to be a viable technique in the debug of advanced microprocessors fabricated on 130 nm as well as 90 nm generation technologies and we believe the technique will scale beyond the 90 nm technology node without significant modifications.

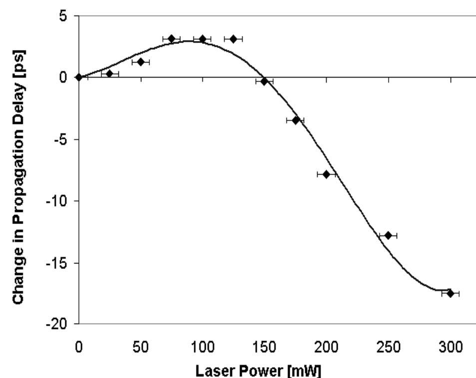

Figure 17. Changes in rising edge delay vs. laser power for a negative edge skewed inverter, i.e., the NMOS device is designed to dominate the logic transition.

VII. Acknowledgments

The authors would like to acknowledge the many technical contributions in the development of LADA by Steve Seidel and also by Jean Hsu both of Intel and the close partnership with Guoqing Xiao, Horst Groneberg, Carlos Manzanilla and Jianxun Mou all of Checkpoint Technologies, LLC (San Jose, CA). Finally, we thank all of the reviewers for their careful reading of the paper and for their many excellent comments.

VIII. References

[1] D. Josephson, S. Poehlman, V. Govan. "Debug Methodology for the McKinley Processor" *Proc. IEEE International. Test Conference*, 2001, pp. 451-460.

[2] T. Eiles, K. Wilsher, W. Lo, G. Xiao. "Optical Interferometric Probing of Advanced Microprocessors". *Proc. IEEE Int. Test Conf.*, 2000, pp. 80-84.

[3] M. Paniccia, T. Eiles, V. Rao, W. Yee. "Novel Optical Probing Technique for Flip Chip Packaged Microprocessors" *Proc. IEEE Int. Test Conf.*, 1998, pp. 740-747.

[4] E. Varner, C. Young, H. Ng, S. Maher, T. Eiles, B. Lee. "Single Element Time Resolved Emission Probing for Practical Microprocessor Diagnostic Applications". *Proc. 28th International Symposium for Testing & Failure Analysis*, 2002, pp. 451-460.

[5] J. Tsang, J. Kash, D. Vallett. "Picosecond Imaging Circuit Analysis" *IBM J. Res. Develop*., 44(4), 2000 p.583-603.

[6] R. Livengood, P. Winer, V. Rao. "Application of Advanced Micromachining Techniques for the Characterization and Debug of High Performance Microprocessors". *J. Vac. Sci. Tech. B*, 17 (1), 1999, pp. 40-43.

[7] R. Guo and S. Venkataraman. "A Technique for Fault Diagnosis of Defects in Scan Chains". *Proc. IEEE Int. Test Conf.*, 2001, pp. 268-275.

[8] K. Cheng "Partial Scan Designs Without Using a Separate Scan Clock". *IEEE 1995*, pp. 277-280.

[9] S. Rusu and S. Tam. "Clock Generation and Distribution for the First IA-64 Microprocessor". *IEEE Int. Sol. State Circuits. Conf.,* 2000, pp. 176-177.

[10] S. Tam, S. Rusu, U. Nagarji Desai, R. Kim, Ji Zhang, I. Young. "Clock Generation and Distribution for the First IA-64 Microprocessor". *IEEE J. Sol. State Circuits* (35), 2000, pp. 1545 – 1552.

[11] M. Bruce, V. Bruce, D. Eppes, J. Wilcox, E. Cole, P. Tangyunyong, C. Hawkins, "Soft Defect Localization (SDL) on ICs". *Proc.28th International Symposium for Testing & Failure Analysis*, 2002, pp. 21-27.

[12] E. Cole. P. Tangyunyong, C. Hawkins, M. Bruce, V. Bruce, R. Ring, W. Chong. "Resistive Interconnect Localization" *Proc.27th International Symposium for Testing & Failure Analysis*, 2001, pp. 43-50.

[13] M. Pronobis and D. Burns. *Proc. International. Symposium for Testing & Failure Analysis*, 1982, pp. 178-181.

[14] D. Burns, M. Pronobis, C. Eldering, R. Hillman. "Reliability/Design Assessment by Internal-Node Timing-Margin Analysis Using Laser Photocurrent-Injection". *Proc. IEEE Int. Reliability Physics Symposium*, 1984 pp. 76-82.

[15] H. K. Brown, G. C. Fuller, M. S. Clamme. "Timing Margin Examination Using Laser Probing Technique". *IEEE* 1990, pp. 384-388.

[16] Y. Taur and T. Ning. Fundamentals of Modern VLSI Devices. Cambridge University Press1998. pp. 129-130.

[17] E. Cole, P. Tangyunyong, D. Barton. "Backside Localization of Open and Shorted Interconnections". *Proc. IEEE Int. Reliability Physics Symposium.* 1998 pp. 129-136.

[18] S. Kolachina, B. Taylor, K. Wills, E. Cole. "Application of TIVA in Design Debug". *Proc. 26th International. Symposium for Testing & Failure Analysis*, 2000, pp. 497-501

[19] S. Thompson, N. Anand, M. Armstrong, C. Auth, B. Arcot, M. Alavi, P. Bai, J. Bielefeld, R. Bigwoood, J. Brandenburg, M. Buehler, S. Cea, V. Chikarmane, C. Choi, R. Frankovic, T. Ghani, G. Glass, W. Han, T. Hoffmann, M. Hussein, P. Jacob, A. Jain, C. Jan, S. Joshi, C. Kenyon, J. Sivakumar, M. Taylor, B. Tufts, C. Wallace, P. Wang, C. Weber, M. Bohr. "A 90nm Logic Technology Featuring 50nm Strained Silicon Channel Transistors, 7 Layers of Cu Interconnects, Low k ILD, and 1um^2 SRAM Cell". *IEEE Int. Electron. Device Meet.* 2002, pp. 61-64.

[20] T. Eiles and P. Pardy "Liquid Immersion Objective for High-Resolution Optical Probing of Advanced Microprocessors". *Proc.27th International. Symposium for Testing & Failure Analysis*, 2001, pp. 167.