

LASER-ASSISTED DEVICE ALTERATION USING SYNCHRONIZED LASER PULSES

Cross-reference to related applications

[0001] This Application is a Continuation-in-Part of U.S. Patent Application No. 13/228,369, filed on September 8, 2011, which claims priority benefit from U.S. Provisional Application Number 61/381,023, filed on September 8, 2010, the disclosures of which are incorporated herein by reference in their entirety. This Application also claims priority benefit from U.S. Provisional Application Number 61/648,042, filed on May 16, 2012, the disclosure of which is incorporated herein in its entirety.

Government License Rights

[0002] This invention was made with government support under USAF Contract FA8650-11-C-7104 awarded by IARPA. The government has certain rights in the invention.

Background

1. Field of Invention

[0003] This invention is in the field of laser-based defect localization analysis of integrated circuits (IC). More specifically, this invention is about design debug and/or failure analysis of ICs using the laser assisted device alteration (LADA) technique.

2. Related Art

[0004] LADA (Laser-Assisted Device Alteration) is a technique that depends on the ability of a continuous-wave (CW) laser to generate localized photocurrents in an integrated circuit through its backside and thus change the pass/fail outcome of a test stimulus on a “sensitive” transistor, thereby localizing sensitive areas including design or process defects. The laser is used to temporarily alter the operating characteristics of transistors on the device. The current spatial resolution using the 1064nm continuous wave laser is 240nm.

[0005] An explanation of the LADA techniques can be found in, for example, *Critical Timing Analysis in Microprocessors Using Near-IR Laser Assisted Device Alteration (LADA)*,

Jeremy A. Rowlette and Travis M. Eiles, International Test Conference, IEEE Paper 10.4, pp. 264-273, 2003. That paper described the possibility of using a CW laser of 1064nm or 1340nm wavelength. It is explained that the 1340nm would cause device alteration via localized heating, while the 1064nm would cause device alteration via photocurrent generation. It is also noted that the 1064nm laser has a spatial resolution advantage. Therefore, the authors recommend the use of 1064nm laser.

[0006] As shown in Figure 1, conventional LADA uses a CW laser to induce electron-hole pairs in the device under test from the backside. The electron-hole pairs so generated affect the timing of the nearby transistor—thus facilitating critical path analysis. A DUT 110 is coupled to a tester 115, e.g., a conventional Automated Testing Equipment (ATE), which is connected to computer 150. The ATE is used in a conventional manner to stimulate the DUT with test vectors and study the response of the DUT to the test vectors. The response of the DUT to the test vectors can be further investigated using the LADA. For example, if the DUT fails a certain test, LADA can be used to investigate whether the DUT can pass under certain conditions and, if so, which device, i.e., transistor, was responsible for the failure. Conversely, when the DUT passes certain tests, LADA can be used to investigate under which conditions the DUT will fail these tests and, if so, which device, i.e., transistor, was responsible for the failure.

[0007] The LADA system for Figure 1 operates as follows. Tilttable mirrors 130 and 135 and objective lens 140 are used to focus and scan a beam from CW laser 120 onto the DUT 110. This allows the laser 120 to generate photo carriers in the silicon of the DUT without resulting in localized heating of the device. The electron-hole pairs so generated affect the timing of the nearby transistor, i.e., decreasing or increasing transistor switching time. The tester is configured to place the operating point of the device under test in a marginal state by applying a recurrent test loop of selected voltage and frequency. The laser stimulation is then used to change the outcome of the tester's pass/fail status. The beam's location at each point is correlated to the pass/fail outcome of the tester, so that when a change is detected, i.e., a previously passing transistor is now failing or vice versa, the coordinates of the laser beam at that time points to the location of the "borderline" transistor.

[0008] During the LADA analysis, the tester (ATE) is configured to place the operating point of the device under test in a marginal state. The laser stimulation is used to change the outcome of the tester's pass/fail status. The present state-of-the-art in laser assisted fault spatial localization is about 240nm resolution. The limitation on further improvement of the single photon LADA spatial resolution is imposed by the laser light wavelength. As noted in the Rowlette paper, the spatial resolution is enhanced by using shorter wavelength. However, optical absorption of silicon at wavelength smaller than 1064nm prevents the use of shorter wavelengths, as it becomes the major obstacle for delivering light to the transistor through the backside. Thus, while design rules of modern devices shrink, the spatial resolution of the LADA system cannot be improved by the use of smaller wavelength laser. For example, at 22nm design rule it is doubtful that conventional LADA will be able to resolve among four neighboring transistors.

[0009] Optical beam induced current (OBIC) is another test and debug analysis in which laser beam is illuminating the DUT. However, unlike LADA, OBIC is a static test, meaning no stimulus signal is applied to the DUT. Instead, the laser beam is used to induce current in the DUT, which is then measured using low-noise, high-gain voltage or current amplifiers. OBIC has been used in a single-photon mode and in a two-photon absorption mode, sometimes referred to as TOBIC or 2P-OBIC (two-photon optical beam induced current).

[0010] Two-photon absorption (TPA) is the simultaneous absorption of two photons of identical or different frequencies in order to excite a molecule from one state (usually the ground state) to a higher energy electronic state. The wavelength is chosen such that the sum of the photon energy of two photons arriving at the same time is equal to the energy difference between the involved lower and upper states of the molecule. Two-photon absorption is a second-order process, several orders of magnitude weaker than linear (single-photon) absorption. It differs from linear absorption in that the strength of absorption depends on the square of the light intensity, thus it is a nonlinear optical process.

SUMMARY

[0011] The following summary of the disclosure is included in order to provide a basic understanding of some aspects and features of the invention. This summary is not an extensive

overview of the invention and as such it is not intended to particularly identify key or critical elements of the invention or to delineate the scope of the invention. Its sole purpose is to present some concepts of the invention in a simplified form as a prelude to the more detailed description that is presented below.

[0012] Various embodiments disclosed enable higher spatial resolution of fault localization by utilizing time domain to enable enhanced spatial resolution. Disclosed embodiments utilize a pulsed laser with sufficient energy in lieu of a continuous wave laser. The pulsed laser is synchronized to the device's clock, thus enabling improved spatial resolution. Various embodiments utilize a 1064nm wavelength laser for a single-photon LADA, or longer wavelengths so as to exploit the non-linear two-photon absorption mechanism to induce LADA effects. This later technique is referred to herein as Two-Photon Laser Assisted Device Alteration technique (2pLADA).

[0013] Disclosed embodiments enable higher resolution of fault localization by using test vectors stimulating a DUT and at the same time use a femtosecond pulsed laser to scan an area of interest in the DUT and examine the response of the DUT to the test vectors during the scan. The laser source is chosen such that the wavelength provides photon energy below the band gap of silicon and it provides pulses of femtosecond pulse width. A clock signal is obtained from the ATE and is fed to the DUT and to the circuit controlling the pulsed laser. The timing of the pulses can be shifted relating to the ATE clock, so as to investigate the pass/fail characteristics of various devices. Additionally, by using proper synchronization of the laser pulses to the clock, spatial resolution is enhanced to enable resolving multiple devices, i.e., transistors, within the laser beam.

[0014] In alternative embodiments, a fixed pulse laser system is used. The clock of the fixed pulse laser is sent to the ATE and is used to derive the test signals to the DUT. Moreover, in order to achieve the effect of shifting the timing of the pulses relative to the test signal, the clock signal of the fixed pulsed laser is shifted prior to sending it to the ATE. Consequently, as the ATE generates the test signals based on the shifted laser clock, the test signal itself is shifted relative to the laser pulses. By controlling the shift of the laser clock, the test signals are shifted relative to the laser pulses to enable the LADA investigation.

[0015] Various embodiments provide a laser assisted device alteration (LADA) system operable in conjunction with an automated testing equipment (ATE) for testing integrated circuit device under test (DUT), comprising: a controller receiving and analyzing test signals from the ATE; timing electronics receiving a clock signal from the ATE, the timing electronics including a first feedback loop generating a synch signal for synchronization of laser pulses to the clock signal; a tunable pulsed laser source generating the laser pulses and having a tunable laser cavity and having a second feedback loop to control the tunable laser cavity so as to generate desired pulse rate of the laser pulses; optical arrangement receiving laser pulses from the tunable pulsed laser source and directing the laser pulses onto desired location on the DUT; wherein the timing electronics is configured for timing the laser pulses to arrive at transistors in the DUT at times synchronized to the clock time to thereby alter the transistors response to test signals applied to the DUT from the ATE, and wherein the controller is configured to detect the altered transistors response. The first feedback loop and/ or the second feedback loop may comprise phase-locked loop. The optical arrangement may comprise a laser scanning microscope. The optical arrangement may further comprise a solid immersion lens. The pulse rate of the laser pulses may be configured as a multiple of the clock signal. The multiple may be an integer larger than one or a fraction. The timing electronics may be configured to delay or advance the laser pulses with respect to the clock signal.

[0016] According to further embodiments, a laser assisted device alteration (LADA) system operable in conjunction with an automated testing equipment (ATE) for testing integrated circuit device under test (DUT) is provided, comprising: a controller receiving and analyzing test signals from the ATE; a fixed pulsed laser source generating laser pulses and generate pulse rate signal indicative of pulse rate of the laser pulses; timing electronics receiving the pulse rate signal and sending a clock signal to the ATE; optical arrangement receiving laser pulses from the fixed pulsed laser source and directing the laser pulses onto desired location on the DUT; wherein the timing electronics is configured for adjusting the timing test signals output by the ATE to arrive at transistors in the DUT at times synchronized to the laser pulses, to thereby detect whether the laser pulses alter the transistors response to the test signals applied to the DUT from the ATE, and wherein the controller is configured to detect the altered transistors response.

[0017] The timing electronics may further comprise a variable phase circuit configured to vary the phase of the clock signal with respect to the pulse rate signal. The variable phase circuit may be configured to retard or advance the test signals with respect to the pulse rate signal.

[0018] According to yet further embodiments, a method for testing an integrated device under test (DUT) that is coupled to an automated testing equipment (ATE) using laser assisted device alteration (LADA) is provided, the method comprising: obtaining clock signal from the ATE and applying the clock signal to the DUT; obtaining a test loop signal and applying the test loop signal to the DUT; applying a first feedback loop to a pulsed laser source to generate laser pulses at repeatable rate; applying the clock signal to a second feedback loop to synchronize the laser pulses to the clock signal; and directing the laser pulses onto desired areas of the DUT. The first and or second feedback loops may comprise a phase locked loop having an external reference signal. The external reference signal may comprise the clock signal. The laser pulses may comprise picosecond to femtosecond laser pulses. The pulsed laser source may be operated to generate laser pulses having wavelength selected to generate single-photon laser assisted device alteration. The pulsed laser source may be operated to generate laser pulses having wavelength selected to generate two-photon laser assisted device alteration. Directing the laser pulses may comprise scanning the laser pulses over a desired area of the DUT.

[0019] According to other embodiments, a method for testing an integrated device under test (DUT) that is coupled to an automated testing equipment (ATE) using laser assisted device alteration (LADA) is provided, the method comprising: using a fixed-pulse laser source to generate laser pulses at a designated pulse rate; obtaining pulse rate signal from the fixed-pulse laser source and generating clock signal therefrom; applying the clock signal to the ATE and generating a test loop signal by the ATE and applying the test loop signal to the DUT; and directing the laser pulses onto desired areas of the DUT. The method may further comprise varying the phase of the clock signal with respect to the pulse rate signal.

[0020] According to further embodiments, a method for measuring carrier lifetime in a semiconductor device under test (DUT) using laser assisted device alteration (LADA) technique is provided, comprising: repeatedly applying electrical test signal to the DUT; using a laser beam

to illuminate a sensitive transistor in said DUT, so as to obtain an optimum LADA signal, thereby obtaining spatial coordinates of the sensitive transistor; using a pulsed laser source to generate laser pulses and illuminating the sensitive transistor using the spatial coordinates to direct the laser pulses; varying timing of the laser pulses and recording strength of the LADA signal for each timing position until the strength of the LADA signal has reached zero; and determining the carrier lifetime by calculating a time period taken for the LADA signal to change from maximum to minimum signal. The laser beam for determining the spatial coordinates may be obtained from a continuous-wave (CW) laser source. The step of calculating the time period may comprise plotting the LADA signal strength vs. time. The method may further comprise a step of directing the laser pulses at the spatial coordinates and varying the timing until optimum LADA signal strength is obtained, so as to determine the pulse timing for maximum laser pulse overlap with test signal overlap.

BRIEF DESCRIPTION OF DRAWINGS

[0021] Other aspects and features of the invention would be apparent from the detailed description, which is made with reference to the following drawings. It should be appreciated that the detailed description and the drawings provides various non-limiting examples of various embodiments of the invention, which is defined by the appended claims.

[0022] The accompanying drawings, which are incorporated in and constitute a part of this specification, exemplify the embodiments of the present invention and, together with the description, serve to explain and illustrate principles of the invention. The drawings are intended to illustrate major features of the exemplary embodiments in a diagrammatic manner. The drawings are not intended to depict every feature of actual embodiments nor relative dimensions of the depicted elements, and are not drawn to scale.

[0023] Fig. 1 illustrates a prior art CW LADA system.

[0024] Fig. 2 illustrates an embodiment of pulsed-laser LADA system.

[0025] Fig. 2A illustrates an embodiment of the two feedback loops.

[0026] Fig. 2B illustrates an embodiment using a fixed-pulsed laser source to generate clock signal.

[0027] Fig. 3 illustrates an embodiment to achieve a synchronization scheme.

[0028] Figs. 4A-4C illustrate how pulsed laser LADA helps individually identify and isolate both the P- and NMOS transistor locations in close proximity.

[0029] Figs. 5A-5C illustrate enhanced spatial resolution as a consequence of accurate pulse placement capabilities.

[0030] Fig. 6 illustrates an embodiment of laser repetition rate locking scheme.

[0031] Figs. 7A-7C illustrate steps taken to measure minority carrier lifetime using LADA, according to one embodiment.

DETAILED DESCRIPTION

[0032] Figure 2 illustrates an embodiment which uses a pulsed laser source with sufficient energy in lieu of a continuous wave laser. This embodiment concerns applying photon absorption to precisely inject carriers into an IC for the purposes of fault localization using the LADA technique and can be used for IC characterization and finding ways to improve the design. The technique is based on photons arriving at a focus in the transistor so that the delivered photon energy is greater than what is needed for electron hole pair creation ($>1.1\text{eV}$ in silicon, for example; other ICs such as GaAs, SiGe, InP, etc, have different bandgap energies). Photon stimulation in this embodiment requires excitation with laser pulses in the range of nanoseconds to femtoseconds. The signal is localized to the focal spot of the laser, providing an immediate improvement in fault localization. The effective volume in which the electron-hole pair generation occurs seems to be reduced due to the synchronization. The embodiments use sophisticated timing electronics to precisely control the timing of the laser pulse with respect to the transition of an edge of the tester clock, e.g., ATE clock. This type of control allows to finely vary the delay or advance of signals propagating through the transistor of interest for LADA.

[0033] Figure 2 illustrates an embodiment of the invention, wherein a DUT 210 is coupled to an ATE 215, as in the prior art. However, in the embodiment of Figure 2, nanosecond to femtosecond laser pulses are generated by pulsed laser source 225, which are then focused onto the DUT 210 using tiltable mirrors 230 and 235 and objective lens 240. For 2pLADA the laser source 225 provides a pulsed laser beam of wavelength that is longer than the silicon bandgap, i.e., longer than 1107nm. In one embodiment wavelength of 1550 nm is used, while in another 1340nm or 1250nm are used. On the other hand, the same arrangement can also be used for single-photon LADA, in which case the laser source can provide pulsed beam of wavelength such as 1064nm. In this embodiment, the tiltable mirrors 230 and 235 are implemented as a laser scanning microscope (LSM). Also, in some embodiments a solid immersion lens (SIL) is used as part of the objective lens arrangement 240.

[0034] In conventional LADA systems the laser is always on; however, according to embodiments of the invention, very short pulses are used. Therefore, it is important that the device transition occurs when the laser pulse arrives at the device. To achieve that, a trigger signal 245 is obtained from the ATE and input to timing electronics 260, which controls the pulsed laser 225 to synchronize the laser pulses with the test signals of the ATE.

[0035] Using the system shown in Figure 2, first the tester (ATE) 215 is operated to apply a set of test vectors to determine the marginal settings of the DUT 210. That is, the voltage and frequency of the test vectors are varied to determine the point where the DUT is just about to fail, or has just failed the test, e.g., the DUT is failing the test 50% of the times of a test loop. This is the DUT's pass/fail boundary condition. The voltage and frequency settings are then used to generate a repetitive test signal to repeatedly stimulate the DUT at its pass/fail boundary condition.

[0036] As the DUT is stimulated at the boundary condition, a sync signal 245 is sent from the tester 215 to the timing electronics 260. The timing electronics 260 controls the laser source 225 to obtain laser pulses of picosecond to femtosecond pulse width and of wavelength higher than silicon band gap for 2pLADA or shorter for single-photon LADA. In general, for 2pLADA the wavelength is about 1250nm to 1550nm and the pulse width is about 100fs. For single-photon LADA the wavelength may be 1064nm and the pulse width is about 100fs. The

laser pulses are scanned over an area of interest in the DUT 240 to thereby increase or decrease the DUT's switching time and push the DUT beyond the boundary condition. That is, if the voltage/frequency of the test vector are set such that the DUT is just about to fail, the laser pulses are timed to cause the DUT to fail. Conversely, if the voltage/frequency of the test vector are set such that the DUT is just failing, the laser pulses are timed to cause the DUT to pass the test. During this time the output of the DUT is monitored to determine location of the failure. That is, at the moment in time where the output signal from the DUT indicates a failure (where without the laser beam the DUT would pass), the location of the beam over the DUT is determined, to thereby determine the location of the transistor causing the failure. Conversely, at the moment in time where the output signal from the DUT indicates a pass (where without the laser beam the DUT would fail), the location of the beam over the DUT is determined, to thereby determine the location of the transistor previously causing the failure and now passing.

[0037] It should be appreciated that since a sync signal is obtained from the tester, the timing of the laser pulses can be varied so as to vary the amount of the photo-generation (single-photon or two-photon) effect on the transistor. That is, the timing of the laser pulses can be varied so as to increase or decrease the amount of increase or decrease the DUT's switching time. This ability can assist in determining the severity of the fault, in addition to its location.

[0038] Embodiments of the invention also use timing electronics to precisely control the timing of the laser pulse with respect to the transition of an edge of the tester (e.g., ATE) clock. This type of control allows to finely vary the delay or advance of signals propagating through the transistor of interest. In one example, two phase locked loops (PLL) are used to precisely control the pulsed laser, as illustrated in Figure 2A. In Figure 2A the ATE 215 provides a clock signal, Clk, and a test loop signal, Test Loop. Both clock signal and test loop signal are input to the DUT and are tapped and sent to the timing electronics 260, which forms the first PLL. The laser source 225 includes a second PLL.

[0039] That is, the PLL of the laser source 225 ensures that the pulse frequency of the laser pulses is stable and accurate to the desired frequency, e.g., 100MHz. Conversely, the first PLL of the timing electronics provides synchronization of the frequency of the second PLL to the clock signal of the ATE. Notably, in this context, synchronization does not necessarily mean

that the laser pulses and the clock pulses are concurrent, but rather that they are synchronized over a test loop period. So, for example, the timing of the laser pulses may be shifted, such that the pulses appear at the middle of the clock signal for every clock pulse, as illustrated by pulse train 227, or at the end of each clock pulse, as shown by pulse train 229, etc. That is, the laser pulses may be delayed or advanced with respect to the clock signal of the ATE, but remain synchronized to the clock signal of the ATE.

[0040] On the other hand, as will be elaborated below, the frequency of the laser pulses may be a multiple of the ATE clock signal. For example, laser pulse train 223 has a multiplier of 7, such that seven laser pulses are generated for every one clock pulse of the ATE. Using a multiplier larger than 1, one can probe whether the failure is at the rising edge, trailing edge, etc. Also, it is not necessary to provide delay or shift of the pulses, since the plurality of laser pulses for each clock pulse serve the advance/delay function. Conversely, it is also possible to have a multiplier smaller than 1. For example, in the pulse train 224 the multiplier is $\frac{1}{2}$, such that a laser pulse arrives only every other clock signal. Such an arrangement can be used to validate that the failure is indeed due to the laser pulses, since if it is due to the laser pulses, then the device would fail about 50% of the time.

[0041] Figure 3 illustrates an embodiment to achieve a synchronization scheme. The output pulses from a pulsed laser source (1) of nano-to-femtosecond temporal duration can be synchronized to the clock cycle of an integrated circuit (IC) (2) through an intermediary phase-locked loop (PLL) circuit (3). In this configuration, the PLL circuit accepts the clock cycle frequency of the IC and locks it to an internal crystal oscillator of the same frequency. In this embodiment, the clock and crystal oscillator frequencies are fixed at 100MHz. The IC clock signal can be generated by the ATE (not shown). This enables a 1:1 optical pulse to transistor-switching-event synchronization ratio. Under these conditions, practically speaking, these values can be fixed anywhere in the range from 1 kHz to 10 GHz, before the efficiency of the photon absorption rate becomes severally diminished.

[0042] It is noted here that optical sources faster than 1GHz are not desirable for nonlinear studies such as with 2pLADA, since in general the peak optical power contained with each pulse is inversely proportional to the repetition rate. Therefore, high repetition rates equal

low peak optical powers which results in ineffective, if any, two-photon absorption. For 1:1 synchronization ratio single-photon studies at 1064nm on the other hand, multi-GHz optical sources may be beneficial since the optoelectronic interaction scales linearly with incident optical power. In addition, it should be noted that the efficiency of two-photon absorption is directly proportional on the incident pulse duration, where femtosecond optical pulses facilitate higher peak optical powers than a picosecond or nanosecond alternative and therefore providing improved nonlinear absorption. As a result, it is desirable for nonlinear studies to utilize ultrafast (i.e. picosecond or femtosecond) optical pulses. On the other hand, for single-photon studies, the pulse duration is not a limiting parameter with regard to absorption rate and so does not restrict performance. If anything, it enables an additional probing parameter (e.g. for measuring optical pulse interaction periods vs. photoelectric device stimulation). Furthermore, the silicon absorption coefficient is greater for single-photon wavelengths < 1130 nm compared to two-photon-absorption-tuned wavelengths (i.e. $\lambda \geq 1250$ nm).

[0043] In order to maintain the efficiency, one can modify the synchronization scheme in order to match integer multiples of incident optical pulses to transistor switching events (or device clock frequencies). For this to succeed, the laser source must be designed to produce a higher than 1 GHz repetition rate and have a scalable pulse picker module in place, post pulse optimization, in order to amend the synchronization ratios. For example, instead of matching each incident optical pulse to every transistor switching event, one could match every second pulse to every subsequent switching event; thus, creating a 2:1 synchronization ratio. Practically, this would translate into the use of a 200MHz repetition rate laser and a 100MHz device frequency, or a 1GHz repetition rate laser and a 500MHz device frequency, and so on. Alternatively, one could tune the ratio to be 3:1 or 4:1, etc., so long as the ratio corresponds to an optoelectronic overlap with the clock pulse in synchronization with the test loop signal. Under this synchronization scheme, the efficiency of photon absorption is not decreased; however, the rate at which absorption takes place does, generating a photon signal intensity that is negatively scaled as per the synchronization ratio. It must be noted here that this may not be a limiting factor for laser-induced studies of integrated-circuits. For each device under test, there is a requirement to perform a photon scalability calibration in order to determine the maximum synchronization ratio allowed. Furthermore, by integrating a tunable optical source (i.e. 1000 –

1600nm output wavelength) into such a system, one can interchange between single-photon and two-photon absorption regimes since two-photon absorption begins to dominate over single-photon absorption at wavelengths larger than about 1200nm in silicon.

[0044] Once these frequencies (i.e. the clock and crystal oscillator frequencies) are locked together, the PLL circuit output signal is then sent through a 100MHz (or clock frequency) electronic filter to the pulsed laser to serve as its input stimulus. The benefit here is that the PLL circuit has full control over the phase of its output signal. Therefore, one can control the repetition rate, and hence the pulse arrival time, of the laser's optical output. This can be verified by comparing the output clock frequency from the IC to the trigger output from the pulsed source on an oscilloscope (9). In this example the PLL circuit can electronically accommodate phase delays of about 600fs; however, due to the board's electrical jitter, a minimum phase delay is set to about 20ps. The system's electrical jitter is directly proportional to the accuracy at which the optical pulses can be positioned relative to the switching time of the individual transistors under evaluation. Therefore, since the system's electrical jitter is 20ps, the optical placement accuracy is also 20ps – there is a one-to-one correspondence. This is a crucial parameter since it can negate the timing benefits to be gleaned if the electronic placement error is larger, for instance, than the 2pLADA femtosecond pulse duration. Femtosecond optical pulses increase the local energy density as required for efficient two-photon absorption; however, when the electrical jitter eclipses this isolated carrier generation timescale then the jitter can limit the subsequent signal generation and temporal precision of the now time-resolved data.

[0045] The laser pulses are then coupled into a laser scanning microscope (LSM) (4) where they can be accurately distributed across a particular location on the IC. The LSM is controlled using a computer (6) which hosts a graphical user interface as well as a custom digital signal processor (DSP) suite. In disclosed embodiments, this applications suite gives the end user, through a preconfigured DSP circuit (7), the ability to directly communicate with the PLL circuit which in turn provides full control over the laser pulse's arrival time at the device, e.g., by delaying or advancing the pulses.

[0046] With regard to the device 2, it can be electrically stimulated to produce preconditioned LADA pass/fail values through the use of a custom applications interface (5). This board compares a real-time acquisition value from a counter, latch and shift register arrangement against a loaded reference value which was inserted beforehand by selecting a reset switch. Fine control over the real-time loaded counter value can be controlled through an analog fine delay potentiometer on the applications interface board which alters the timing of the latch enable function on the IC. This configuration allows the user to condition predominantly passing, failing or balanced comparator output values. These pass/fail output values are then sent to a data conditioning circuit (in this example, a field-programmable gate array FPGA 8) which is programmed to accept a real-time digital pass/fail stimulus, conditioned to scale between failing values from 0-100%, and delivers an averaged (about 40us period) digital output, again scaled between 0-100% fail, for enhanced visualization and biasing of the resulting pass/fail levels in the graphical user interface. The data conditioning circuit can also be used in conjunction with the applications interface board in order to calculate the magnitude of laser-induced timing delays by calibrating the application board's output delay voltage.

[0047] In the embodiments described above, a tunable pulsed laser source is used, and the pulse frequency is tuned so as to synchronize to the ATE clock. While these embodiments are operable, a tuned pulsed laser source is rather expensive and requires the phase locked loop described above. Figure 2B illustrates another embodiment, which enables LADA testing using a simpler fixed pulsed laser 255. For example, a mode-locked laser source can be used. Mode-locking is a technique in optics by which a laser can be made to produce pulses of light of extremely short duration, on the order of picoseconds or femtoseconds. The laser pulses are used as the clock, which is fed to the timing electronics 265. Conventional ATE's have clock input port and are programmable to use the input clock in order to generate the clock, Clk, and the test loop signals for the DUT. Therefore, in one example, the clock signal from the timing electronics 265 is input to the ATE, and the ATE is programmed to use that input to generate the clock and test loop signal.

[0048] However, as noted above, to obtain the most benefit of the pulsed laser LADA, it is desirable to adjust the pulses such that the laser pulses arrive at the transistor at different times

during the clock cycle, e.g., leading edge, middle, trailing edge, etc. In the embodiments of Figure 2, 2A and 3, this was done by advancing or retarding the laser pulses. However, in the embodiment of Figure 2B the laser pulses are fixed and cannot be changed, so retarding or advancing the laser pulses is not a viable option. Therefore, according to one embodiment, the ATE is programmed to retard or advance its clock signal in synchronization with the clock signal received from the timing electronics 265. In this manner, the timing of the arrival of the laser pulses at the transistor can be tuned to leading edge, trailing edge, etc., of the ATE clock signal.

[0049] On the other hand, since in general the ATE and the LADA testers are made by different manufacturers, and the actual testing is performed by an engineer from yet a third company, it would be beneficial to simplify the operations of the testing engineer and offload the retardation or advancement of the signal from the ATE. This is done by using the phase shifter 275, illustrated in the embodiment of Figure 2B. That is, using the phase shifter 275, the clock signal output from the timing electronics 265 can be advanced or retarded with respect to the laser pulses. The resulting modified signal is then sent to the ATE as the input clock signal. Consequently, when the ATE output its clock and test loop signals, they are both can be shifted or retarded with respect to the laser pulses.

Examples

[0050] Constructing a pulsed LADA system with a pulsed optical source allows for novel aspects of the operating device to be evaluated and measured. When traditional single-photon or alternative two-photon LADA utilize a CW laser, the optical radiation is constantly interacting with the individual transistors with a potentially damaging level of invasiveness. A pulsed LADA methodology, on the other hand, allows the switching characteristics of the individual transistors to be mapped out in as much as two-physical dimensions. Extended pulsed LADA concepts are discussed in detail below.

[0051] Under conventional CW LADA stimulation, device theory and practice informs us that the magnitude of laser-induced device perturbations from a p-type metal-oxide-semiconductor (PMOS) transistor dominates over its n-type (NMOS) neighbor. Since the diameter of the laser beam would cover both the p-type and its neighboring n-type transistors, the resulting spatial resolution is insufficient to distinguish the failing transistor. On the other hand,

using the disclosed embodiments in a pulsed arrangement, temporal resolution is used to achieve higher spatial resolution, even when a larger wavelength laser is used. That is, since the incident pulses are tuned to the exact temporal switching interval of the transistors under investigation, and also because the peak power contained in each pulse is significantly higher than in CW mode, it is possible to individually identify and isolate both the P- and NMOS transistor locations in close proximity. This is impossible under CW excitation and therefore creates a new experimental avenue for semiconductor device design debug and characterization to explore, even at increasingly small design rules. This addresses a growing concern within the semiconductor device failure analysis community where optically-induced transistor recognition and operating characteristics are of critical importance as the latest technology nodes are scaled toward lower nanometer geometries. Consequently, synchronized pulsed LADA offers more value than its CW counterpart.

[0052] A schematic example of such improvements is presented in Figs. 4A-4C. In continuous-wave mode, the PMOS signal generally dominates, resulting in a generalized spatial distribution of a single signal, as illustrated in Figure 4A. It is incredibly difficult here to discern individual physical transistor distributions and/or match these LADA manifestations to computer-aided design (CAD) layouts. Theoretically, each transistor should generate its own LADA signal – no matter the magnitude of the laser-induced effect, as illustrated in Figure 4B. These would perfectly trace out the physical location of the individual transistors allowing for rapid physical and/or optoelectronic recognition. This can be replicated in the pulsed domain using the embodiments discussed above. That is, the laser pulses are timed and synchronized to the test signal, such that they arrive at the P- and NMOS transistor locations according to the user's selection. The pulses can be timed to the switching of the PMOS transistors to test them, or to the switching of the NMOS to test the NMOS transistors, as illustrated in Figure 4C. Therefore, unique transistor switching evaluations as well as CAD-enhanced physical mapping/recognition of LADA signals can be performed, regardless of the spatial laser coverage.

[0053] Furthermore, an additional benefit of the increased peak power from ultrafast pulses (aside from the ability to generate LADA signals more efficiently, i.e., acquiring fewer image averages) is the generalized increase/decrease (depending on whether you perturb a P- or

N-MOS transistors) in laser-induced critical timing path perturbations and therefore improved LADA signal collection. Larger incident optical power increases the number of photo-injected carriers within the silicon, which in turn enhances the probability of stimulating optoelectronic fluctuations within the device structures. This leads to superior LADA signal responses which can be more readily measured with a reduced level of invasiveness – pulsed optical sources are actually off for longer than they are on, limiting the opportunity for thermal build up and damage. For example, an ultrafast laser with a pulse duration of 10ps and a repetition rate of 100MHz is off for a duration of 10ns – creating a ratio of 1:1000 (ON:OFF) – thereby providing ample cooling off period. However, it should be noted that it is ultimately the power ratio that causes the heating. For example, a single optical pulse containing 1kJ of incident optical energy would satisfy the above criteria however it also contains enough energy to potentially permanently damage the device through some other thermal or non-thermal optoelectronic mechanisms.

[0054] Also, with the facility to noninvasively inject significant levels of optical power into a specific transistor comes the opportunity to disturb previously overlooked transistor locations. Naturally, a greater level of photo-carrier generation within the vicinity of a given region of interest (which is populated with transistors of varying sensitivity) increases the probability of visualizing a wider range of LADA manifestation sites. These activation regions can be stimulated with approximately 10-100uA of laser-induced photocurrent; however, with an ultrafast laser pulse boasting peak optical powers approaching 10-100kW it may be possible to inject 10-100mA of photocurrent (while maintaining a safe level of invasiveness) within the device which may be enough to perturb ‘healthy’ transistors.

[0055] Efficient two-photon absorption can be obtained in silicon with a focal laser power density of more than 10MW/cm²; however, single-photon values will be approximately 10⁶ times smaller due to its relative absorption cross-section. The level of incident optical power (or local power density) required for efficient, non-invasive photo-carrier injection will decrease as the spatial geometries of the transistors under evaluation also decrease. Also the generation of two-photon absorption is not dependent on a specific power density threshold – two-photon absorption is an instantaneous, quantum mechanically defined, nonlinear process which is

sensitive to the imaginary part of the third-order nonlinear susceptibility (i.e. it exhibits an intensity-squared dependence, not a power density dependence).

[0056] Even though a two-photon wavelength of 1250nm generates effectively 625nm inside the silicon (where the absorption cross-section is higher than 1064nm), the intensity-dependent nature of the absorption process reduces its overall relative rate of absorption. Two-photon absorption is directly proportional to the square of the incident optical intensity. In addition, silicon doping levels also contribute to this discussion – i.e. increased or decreased doping concentrations affect the level of absorbance as a function of wavelength. This single-photon-biased opportunity however enables another novel laser-probing and device characterization platform for enhanced critical timing analysis of racing/switching signal levels within transistors. CW LADA cannot offer this type of interrogation due to limitations from invasiveness (i.e. the laser - is always on) and limited power delivery capabilities. Time-resolved pulsed probing, on the other hand, may allow failure analysts to investigate, for the first time, transistor switching physics on healthy, design-defined nodes as well as the subsequent down-chain device performance/interactions also. In order to effectively implement this type of device characterization, an understanding of the level of incident optical power required plays an important role. Perturbing ‘healthy’ transistors requires high peak powers while facilitating a minimal level of invasiveness. With that said, optimization of the temporal duration of the incident optical pulses is required. Clearly, picosecond pulse durations at 1064nm provide significant levels of incident optical power (hence, photo-carrier generation) at the transistor level since, for example, a 10ps laser pulse at a repetition rate of 100MHz and an average power of 4mW generates a peak power of 4W; however, this may be limited if the laser repetition rate is matched to a faster than 1GHz clock frequency from a device under test. Increasing repetition rates results in decreasing peak powers. Therefore, a more suitable alternative is the use of a femtosecond laser source. The laser repetition rate could be scaled in accordance with the device operation frequency while providing enhanced levels of peak optical power since the pulse duration has reduced by a factor of 1000 – increasing the peak power by the same magnitude (4kW in the example given above). An additional benefit of the femtosecond pulse duration would be the improvements in temporal characterization; however, this is restricted by the magnitude of the synchronization scheme’s electrical jitter – as described previously. Finally,

femtosecond laser pulses provide a reduced level of optical invasiveness compared to picosecond or nano-second pulses, minimizing the potential for laser-induced damage to the device.

[0057] In addition, a pulsed LADA system demonstrates enhanced spatial resolution as a consequence of accurate pulse placement capabilities. Again, when in CW mode, the laser continuously stimulates a specific region of interest while inferring LADA information in real-time. This results in a spatially averaged two-dimensional LADA image since there is no discrimination between the highly ordered sequences of circuit functionality (i.e. propagating signal paths vs. time) where a collective distribution is acquired with a PMOS dominated bias. In pulsed mode however, one can discern between these propagating speed paths with an accuracy of about 20ps, allowing for highly confined LADA signal manifestations which have an enhanced lateral resolution since they individually and temporally address spatially separated neighboring transistors that are not configured to switch until later in the device operational cycle. This enhances the LADA isolation resolution as well as the physical LADA resolution.

[0058] A schematic example is presented in Figures 5A-5D. In continuous-wave mode, since the spatial distribution of LADA signals is time-averaged, the resulting two-dimensional LADA map has a generalized optoelectronic structure which may be biased according to the individual transistors' LADA signal strengths (since PMOS generally dominates over NMOS). These images, illustrated in Figure 5A, suffer from poor spatial resolution and limited CAD overlay capabilities. In pulsed mode however, the LADA image can be enhanced and its spatial resolution improved due to the time-resolved nature of the acquisition process. By individually addressing each transistor as a function of both space and time (and with sufficient incident optical power so as to remove any PMOS biasing effects), neighboring transistor influence is effectively removed from perturbing the LADA acquisition since there is now a tester-driven, transistor-dependent sequence of events controlling the device operation. Each transistor is configured to switch in a systematic, time-dependent order, allowing for the incident optical pulse to directly manipulate and measure each transistor in two physical dimensions (i.e. X and Y) as well as time. As a result, the spatial resolution of the acquired LADA signal is improved and therefore additional, previously unobtainable, device data may be extracted, as illustrated in

the sequence of Figure 5B-5C, showing images taken at different times, each with its temporal and spatial separation.

[0059] Aside from gleaning only LADA-specific data from this technique, it is also possible to determine additional optoelectronic phenomena as well. One example would be the measurement of laser-induced carrier lifetimes. Currently, the carrier lifetime inside specific device locations is tremendously difficult to quantify since it depends on a number of different optoelectronic parameters, such as material composition, dimensions, geometries and electric-field magnitudes and directions, etc. With pulsed LADA however, one may be able to directly measure this electronic timescale through a pseudo pump-probe type technique where the creation of a transistor-specific LADA event is linked to the arrival time of a laser pulse. The measured carrier lifetime may require the consideration (i.e. subtraction) of the system's electronic response time for a more accurate representation.

[0060] When using single-photon LADA, i.e., laser pulses of wavelength 1064nm, the magnitude of the measured LADA effect is directly proportional to the magnitude of the laser-induced photo-current (this is when using linear absorption, as the LADA signal would respond quadratically with a two-photon technique). According to one embodiment, the LADA signal is mapped out as a function of laser pulse arrival time. Then it would be possible to extract the carrier lifetime, since the lifetime will dictate the magnitude of the resulting LADA signal.

[0061] A process for accomplishing this according to one embodiment would be as follows. First, a laser beam (e.g., a CW laser beam of wavelength 1064nm) is parked to illuminate a sensitive transistor, so as to obtain an optimum LADA signal. This is illustrated in Figure 7A. The spatial coordinates of the laser beam at the optimum LADA signal indicates the proper spatial coordinates of the transistor. Then the CW laser source is disabled and the pulsed laser source is activated and the laser pulses are directed to the same spatial coordinates obtained from the CW laser. The laser pulse timing is adjusted to obtain and measure the optimum LADA signal, so as to obtain the proper temporal overlap with the tester (e.g., ATE) pulses arriving at the transistor. This is illustrated in Figure 7B. At this time, the optimized the spatial overlap of the laser spot with the transistor and the temporal overlap of the laser pulses with the test signal have been achieved. Then, the laser pulses arrival time can be adjusted to measure the carrier

lifetime. Specifically, the timing of the laser pulses is then varied and the LADA signal strength is recorded for each timing position (e.g., retarded or advanced amount), until the LADA signal has reached zero. The resulting LADA signal strength vs. time response is then plotted, as illustrated in Figure 7C. The time taken for the LADA signal to change from maximum to minimum signal (or vice versa) corresponds to the measured laser-induced carrier lifetime. The above process is performed while repeatedly applying electrical test signal to the DUT.

Laser Source

[0062] Multi-GHz repetition rate laser sources are readily available and are constructed through careful consideration of their resonator cavity length – i.e. the shorter the oscillating cavity, the higher the repetition rate. Control over the cavity length can be manipulated and locked through the inclusion of a piezo-electric actuator located on the opposite side of an intra-cavity resonator mirror. This is the industry standard technique for repetition rate locking; however, the electronic mixer circuit required to facilitate such a scheme may differ in design and implementation. To properly incorporate the tuned-pulsed laser source into the LADA tester as described in the above embodiments, two feedback loops are required; one to control the repetition rate of the laser pulses and one to synchronize the timing of the pulses to the DUT clock. The first feedback loop, which controls the repetition rate, includes a mixer that compares the free-running repetition rate frequency of the laser to an input clock stimulus in order to produce a high-voltage driven difference signal. The difference signal is input to the piezo-electric transducer to adjust the resonator cavity length, which then adjusts the desired length so that the pulse rate matches the clock input supplied. An example of such a setup is illustrated in Fig. 6. In addition to the circuit described in Fig. 6, a secondary stabilization scheme can be included to continually monitor and correct the output voltage from the proportional-integral amplifier. This ensures that the high voltage amplifier is consistently given the correct input voltage so to provide repetition rate locking stability of a greater period of time – i.e. several days compared to tens of minutes.

[0063] It should be understood that processes and techniques described herein are not inherently related to any particular apparatus and may be implemented by any suitable combination of components. Further, various types of general purpose devices may be used in

accordance with the teachings described herein. It may also prove advantageous to construct specialized apparatus to perform the method steps described herein.

[0064] While this invention has been discussed in terms of exemplary embodiments of specific materials, and specific steps, it should be understood by those skilled in the art that variations of these specific examples may be made and/or used and that such structures and methods will follow from the understanding imparted by the practices described and illustrated as well as the discussions of operations as to facilitate modifications that may be made without departing from the scope of the invention defined by the appended claims.

Claim

The invention claimed is:

1. A method for testing an integrated device under test (DUT) that is coupled to an automated testing equipment (ATE) using laser assisted device alteration (LADA), the method comprising:

obtaining clock signal from the ATE and applying the clock signal to the DUT;

obtaining a test loop signal corresponding to marginal setting of the DUT and applying the test loop signal to the DUT;

applying a first feedback loop to a pulsed laser source to generate laser pulses at repeatable rate;

applying the clock signal to a second feedback loop to synchronize the laser pulses to the clock signal;

directing the laser pulses onto desired areas of the DUT;

collecting electrical signals from the DUT;

detecting whether the laser pulses alter the DUT response to the test loop signal applied to the DUT from the ATE.

2. The method of claim 1, wherein the first feedback loop comprises a phase locked loop having an external reference signal.

3. The method of claim 2, wherein the external reference signal comprises said clock signal.

4. The method of claim 1, wherein the second feedback loop comprises a phase locked loop.

5. The method of claim 1, wherein the laser pulses comprise picosecond to femtosecond laser pulses.

6. The method of claim 1, wherein the pulsed laser source is operated to generate laser pulses having wavelength selected to generate single-photon laser assisted device alteration.

7. The method of claim 1, wherein the pulsed laser source is operated to generate laser pulses having wavelength selected to generate two-photon laser assisted device alteration.

8. The method of claim 1, wherein directing the laser pulses comprises scanning the laser pulses over a desired area of the DUT.

9. The method of claim 1, wherein the test loop signal comprises voltage and frequency configured to cause the DUT to fail the test 50% of the time.

10. The method of claim 1, wherein the marginal setting comprises the DUT failing the test 50% of the time when no laser pulses are directed onto the DUT.

11. The method of claim 1, wherein pulse rate of the laser pulses is configured as a multiple of the clock signal.

12. The method of claim 1, further comprising delaying or advancing the laser pulses with respect to the clock signal.

13. The method of claim 1, further comprising controlling the first feedback loop to adjust the timing of the laser pulses arriving at transistors.

14. The method of claim 1, further comprising varying phase of the clock signal with respect to the repeatable rate.

15. The method of claim 1, further comprising focusing the laser pulses onto desired transistor within the DUT at a photon energy that is greater than energy required to generate an electron-hole pair.

16. The method of claim 1, wherein applying the test loop signal to the DUT comprises repetitively applying the test loop signal to repeatedly stimulate the DUT at its pass/fail boundary condition.

17. The method of claim 1, wherein directing the laser pulses onto desired areas of the DUT comprises directing laser pulses configured to increase the DUT fail rate.

18. The method of claim 1, wherein directing the laser pulses onto desired areas of the DUT comprises directing laser pulses configured to increase the DUT pass rate.

19. The method of claim 1, further comprising identifying location on the DUT where the laser pulses focused to thereby identify location of transistors causing failure of the DUT.

20. The method of claim 1, further comprising varying timing of the laser pulses to thereby vary the DUT switching time.