

Product-representative “at speed” test structures for CMOS characterization

M. B. Ketchen
M. Bhushan

The design of product-representative test structures for measuring and characterizing CMOS circuit performance, power, and variability at speeds characteristic of present-day microprocessors is described. The current use of this set of test structures in the IBM partially depleted silicon-on-insulator CMOS technologies covers diagnostics in early process development, monitoring mature processes in manufacturing, enabling model-to-hardware correlation, and tracking product performance. The designs focus on measuring high-frequency performance early in the product fabrication cycle while minimizing test and data analysis time. The physical layouts are compact, facilitating placement in the chip. A subset of these test structures can be measured at the first metal level, while more complex designs use three or more metal layers. Most designs are compatible with standard in-line parametric test equipment, although a limited number of bench tests continue to play an important role. Differential measurement techniques are key to many of the test structure designs. Hardware data analysis also relies heavily on differencing schemes for relating MOSFET parameters and associated parasitic components to circuit delays in a self-consistent manner.

Introduction

With the continued scaling of CMOS technology, the growing contributions of physical layout style and parasitic capacitance and resistance to circuit switching delays, the increasing variety of MOSFET device options to cover low-power and high-performance applications, systematic and random variations in MOSFET parameters, and silicon-on-insulator floating-body effects, ac performance-based technology assessment has become a key component of technology characterization. Common practice such as basing process decisions on measured dc parameters of single MOSFETs and capacitance measurements of separate structures is inadequate in the presence of systematic and random parameter variations. On the other hand, measurements made under high-frequency switching conditions are truly representative of the behavior observed in a CMOS product. These measurement values, averaged over a large number of nominally identical circuits, can be configured to enable the extraction of average MOSFET parameters and in some cases even their distributions under circuit application conditions. This approach,

augmented with traditional dc characterization, can rapidly provide essential information for technology development, for monitoring the manufacturing line, for model building, and for tuning product performance and power.

We have designed a set of product-representative test structures configured to measure device and circuit properties under conditions characteristic of multi-GHz product operation. These test structures for “at speed” characterization have been implemented in the silicon manufacturing line at IBM for partially depleted silicon-on-insulator (PD-SOI), beginning with the 180-nm-technology node. The expanded suite of test structures in 90-nm- and 65-nm-technology nodes covers both logic and SRAM circuits and uses up to six levels of metal. As a tactical measure, the designs are arranged such that the high-speed activity takes place solely within the test structures, and only a low-speed interface with the measurement equipment is required. Automated in-line measurements can be performed with low-frequency contact probes using standard parametric testers. Test structures for occasional off-line bench tests using

©Copyright 2006 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

0018-8646/06/\$5.00 © 2006 IBM

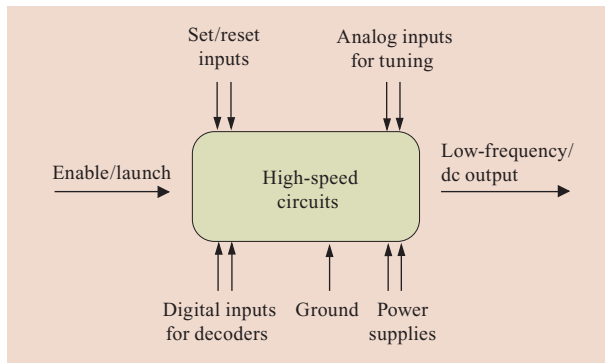


Figure 1

Block diagram of an “at speed” test circuit with dc inputs and low-frequency or dc outputs.

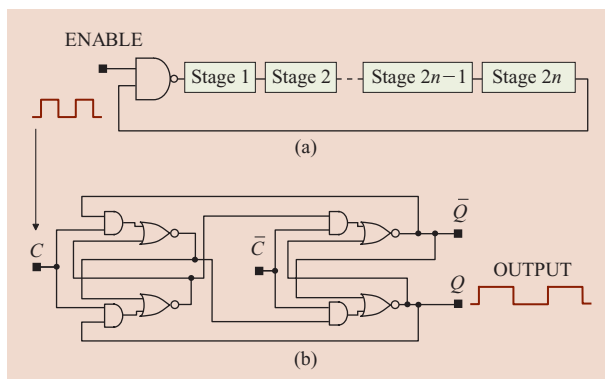


Figure 2

(a) Ring oscillator with $2n$ identical stages and a NAND2 gate to enable the ring; (b) master-slave flip-flop to divide the RO frequency by 2. C , \bar{C} and Q , \bar{Q} are the true and complementary inputs and outputs, respectively.

high-frequency equipment allow a more exhaustive characterization, complementing the in-line tests and providing specific information for model build. In addition, “at speed” test structures embedded in the product and measured after completion of the full chip fabrication process are used to directly connect technology characterization to the product performance. A differencing scheme is frequently used for extracting a single parameter from measurements or analysis of two designs, identical in all respects except for their sensitivities to the selected parameter. Many of the individual test structures themselves are inherently differential in nature, with sources of systematic error and background noise automatically subtracted out. Techniques based on ratios of measured parameters to

their predicted values are used extensively to extract trends and variations. The designs and measurements are geared toward maximizing information content with minimum test time, and cross-correlation among different tests. All of the designs are customized and structured to facilitate direct comparison across different technology generations. Wherever possible, calibration and data analysis are integral parts of the design itself.

In this paper, the basic principles behind the designs of these “at speed” test structures are described. First, the physical structure of “at speed” macros and commonly used circuit building blocks is reviewed. This is followed by a description of in-line test structures incorporating ring oscillators applicable to both bulk silicon and SOI technologies, and in-line test structures specific to PD-SOI technology. In the final section, test structures for more exhaustive characterization using high-frequency bench tests are discussed. Circuit simulation results and examples of data are included as appropriate. A list of common symbols used in the text and figures is given in Appendix A (Figure 19).

Physical structure of macros

The general concept of an “at speed” macro that can be measured using only dc I/Os is shown in **Figure 1**. The primary input signal may enable a ring oscillator or initiate a chain of events in a high-speed digital circuit. The output may be a low-frequency (<5 MHz) signal, which can be measured using an off-the-shelf frequency counter, a dc voltage, or a dc current. The control signals may be digital inputs to decoders to select a circuit under test or to define a path through a digital circuit, digital inputs serving as set or reset controls for latches, and analog inputs to control various voltage or current levels. Additional inputs are power supply, V_{dd} , and ground, GND, connections feeding into a very low-resistance power grid. There are generally two or more independent V_{dd} sectors in a macro that share a common GND.

The in-line test-structure macros are rectangular in shape, with a length of 2.5 mm and a width ranging from 110 μm to 230 μm . All make use of a common 1×25 -pad set with standard $60\text{-}\mu\text{m} \times 90\text{-}\mu\text{m}$ test probe landing pads. Some of the more basic macros are designed to be testable with only the first metal layer, M1, for early learning in the process cycle. More complex macros use up to three or more metal layers. The form factor of the macros is suitable for placement on the kerf (non-product areas at the edges of the chip). However, it also poses design challenges in minimizing voltage drops within the power-distribution system as well as along signal wires traveling across the macro. The circuit schemes utilized in these macro designs are selected for enabling “at speed” tests using only dc I/Os and for compatibility with the physical layout of the macro.

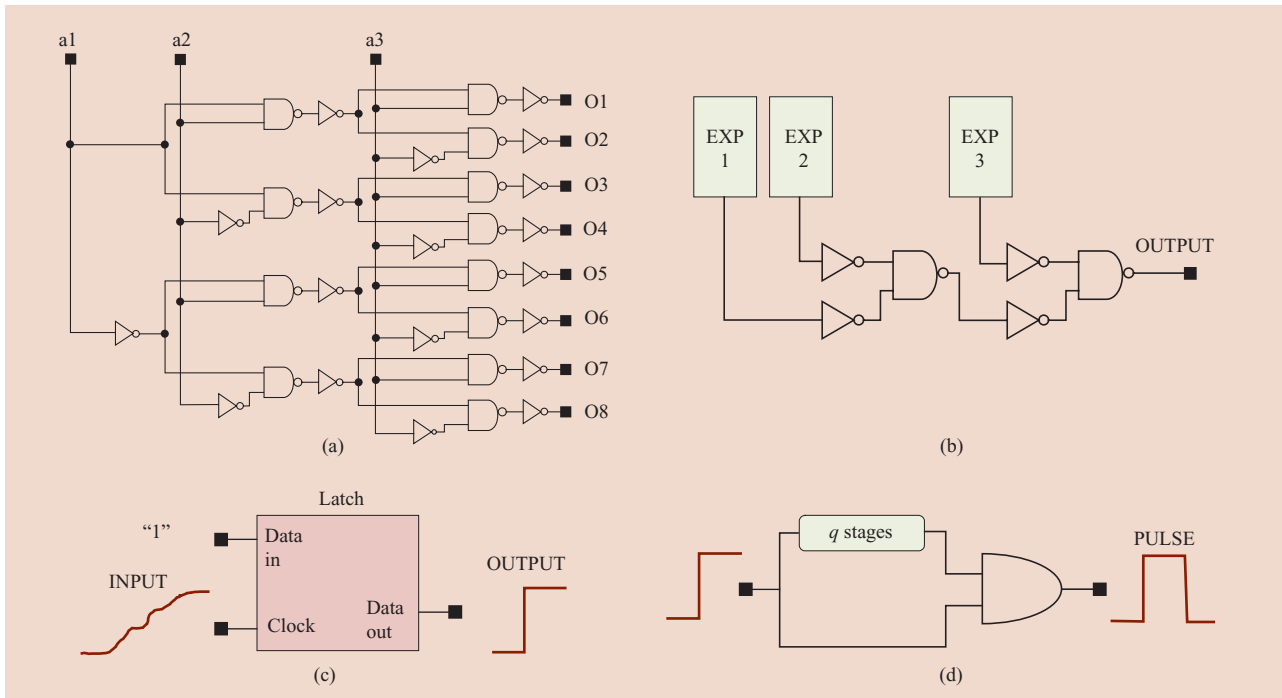


Figure 3

(a) Three-bit decoder circuit with inputs a1, a2, and a3 to initiate or enable one of eight experiments or ROs connected to outputs O1 to O8. (b) “OR” function to drive a common output; the output from each unselected experiment, EXP, is at GND, and OUTPUT follows the output of the selected EXP. (c) Conventional latch for generating a single sharp edge. (d) Circuit using an AND gate to generate a pulse from a sharp edge. Parts (c) and (d) reproduced from [4], with permission; ©2005 IEEE.

Additional test structures that are imbedded directly within a product share the power supply and I/Os with other functions on the chip and have a length-to-width ratio close to unity. There is also a subset of bench-testable designs that have form factors and power distribution arrangements similar to those of the in-line macros. The differential analysis scheme intrinsic to these designs necessitates custom physical layout; however, the layouts are structured in a hierarchical manner that facilitates rapid modifications to accommodate changes in the technology ground rules or migration to another technology node. Different experiments are inserted into a common macro template designed for a specific application, and multiple test structures share the same top-level design and test code.

Test structure building blocks

The macro designs make frequent use of a small number of basic building blocks which are briefly reviewed before discussion of specific macro implementations. These building blocks include ring oscillators, decoders, multiplexers, frequency dividers, latches, and pulse generators. A ring oscillator (RO) shown in **Figure 2(a)** is

commonly employed for circuit delay measurements or for generating a steady pulse stream. The RO comprises $2n$ identical stages and a two-input NAND (NAND2) connected to form a closed loop. One input of the NAND2 is used for changing the state of the RO from a non-oscillating or quiescent state (ENABLE = “0”) to an oscillating one (ENABLE = “1”). The output voltage signal is a square wave when the RO is oscillating and a logical “0” otherwise. The period of oscillation or output frequency is a function of the circuit type and number of stages. This RO frequency is divided by k to lower the output frequency to <5 MHz for external frequency measurements or for use as a clock input to an on-chip digital counter. A circuit for dividing the frequency by two is shown in **Figure 2(b)**. It serves as a stage in a divide-by- k circuit and also as a building block for a counter.

A three-bit decoder circuit, shown in **Figure 3(a)**, sets one of eight possible output signals to a logical “1,” the other seven remaining at “0.” This output can be used, for example, to enable one of eight ROs or experiments. The output signals from multiple ROs or experiments are fed to an OR circuit, as shown in **Figure 3(b)**. The OUTPUT

node follows the output of the experiment which has been enabled, the outputs from all unused experiments being at ground potential. Physical implementations of these two functions with any number of inputs can be local or extended and adapted to a wide variety of experiments within the geometric constraints of the design.

A circuit for generating a sharp edge with a transition time of <20 ps from a slowly rising (dc) voltage input is shown in **Figure 3(c)**. It employs a latch in which a logical “1” is preloaded into the data port, with its output at “0.” As the external dc signal input to the clock of the latch rises, at some point in time the preloaded “1” passes through the latch and emerges as the single sharp edge at the latch output. After this event the latch output remains at “1,” independent of the state of the clock input, until the latch is reset.

A pulse of width calibrated in units of a CMOS gate delay can be generated from a single sharp rising edge using a circuit such as that shown in **Figure 3(d)**. In this case an initial single sharp edge drives both inputs of an AND gate. One of the inputs is delayed by passing it through q inverting gates of a delay chain, where q is an odd number. This results in an output pulse of width qd , where d is the delay per gate. The pulse width can be varied by varying the length of the delay chain. It can also be calibrated by measuring the delay per gate if the same gate design is used to construct a ring oscillator. In some applications, where transient measurements are made on a circuit containing the same gate design, the pulse width becomes self-calibrated.

The latch circuit shown in **Figure 3(c)** can also be used as a pulse detector. With a “1” preloaded into the data port, the latch output transitions from a “0” to “1” if a clock pulse of sufficient strength is applied (i.e., a full rail pulse of width ~ 30 ps or greater). The latch output then remains at “1” and can be observed at a later time.

The logic gates and memory circuits in all of these test structure designs are representative of those used in IBM microprocessors, with a common set of circuits used across multiple test structure designs. Companion test structures with single MOSFETs, identical in physical layout to those in logic and memory circuits, are included for standard dc analysis.

Finally, power distribution is an extremely important consideration in the design of “at speed” macros, especially with the kerf macro form factor and dc I/Os. For a circuit that draws a constant current, such as an RO, the primary consideration is that the dc power droop be kept to a minimum, typically a few mV at most. For designs testable at the first metal level, the power grid for an individual experiment takes the form of interdigitated “fingers” emanating from the sides of adjacent V_{dd} and GND I/O pads or extensions thereof. For more complex macros, a product-like power grid is used with additional

metal strapping wherever possible to further lower the resistance, along with multiple GND and V_{dd} pads as appropriate. For most pulse-based circuits, the current draw is very irregular, and it cannot be assumed that the dc power I/O can provide significant additional charge on the timescale of the experiment. Decoupling capacitance has been included in the macros to ensure that the power-supply voltage droop over the duration of an individual single-shot experiment remains insignificant. In all cases, the I/O driver and other support circuitry are powered by an independent V_{dd} with on-chip decoupling capacitance, and checks are done to ensure that the I/O and support circuitry is not significantly disturbing the circuit under test.

Ring-oscillator-based test structures

Ring oscillators are widely used for measuring circuit-switching delays in the picosecond range [1–3]. The output period is increased to the microsecond range by using a large number of identical circuits in a closed loop, in conjunction with an on-chip frequency divider. For “at speed” characterization of CMOS technology, we have extended the use of ROs to extracting MOSFET device parameters and parasitic components [4]. The basic idea for parameter extraction is to use RO stage designs which, by a differencing technique applied to a pair of ROs, give a measure of a specific critical circuit parameter. For model-to-hardware correlation and in-line characterization, circuit delay tracking methods have been developed for both visual and quantitative data analysis [5]. These techniques are geared toward detecting deviations in the hardware from the predicted circuit behavior of the order of about 3% or more and pointing to the source of the variation. The ROs typically have 100 stages, with two or more MOSFETs in each stage. The measurements are thus averaged over several hundred MOSFETs, which renders them immune to local random variations. Measurements from spatially separated ROs are very useful for mapping systematic variations in parameters across product chips and across entire silicon wafers.

The delay, d , of a stage in an RO shown in **Figure 2(a)** is given by

$$d = \frac{1}{4nkf}, \quad 2n + 1 \gg 1, \quad (1)$$

where f is the frequency measured at the output of the divide-by- k circuit, and each stage switches twice during a complete cycle. If necessary, a more precise determination of the delay per stage can be obtained by using a simulation-based correction for the NAND2 delay. In addition to the RO frequency measurements, the current drawn by the RO power supply during its switching and quiescent states, IDDA and IDDQ respectively, is also

measured. The capacitance per stage of the RO, C_s , is determined from the charge transfer in switching the logic state of each stage,

$$C_s = 2d \frac{IDDA - IDDQ}{V_{dd}}. \quad (2)$$

The delay per stage is also expressed in terms of the switching resistance, R_{sw} , and C_s ,

$$d = R_{sw} C_s. \quad (3)$$

Here $1/R_{sw}$ is a measure of the current drive capability of the logic gate [4, 6].

A set of RO stage designs for extracting resistance and capacitance components of circuits is shown in **Figure 4**. A reference RO stage with an unloaded inverter is shown in Figure 4(a). In Figure 4(b), the output of the inverter is connected to an n-FET whose average gate capacitance over the switching cycle is determined from the difference in RO capacitance from the reference stage in Figure 4(a). In Figure 4(c), a metal wire is added which is configured to add wire capacitance, C_w , where the wire resistance, R_w , is much smaller than that of the inverter, R_{sw} . If R_w is made comparable to R_{sw} , information on wire RC delay is obtained. Figure 4(d) shows an n-passgate circuit consisting of an n-FET pass-transistor, NPG, following an inverter. In this case, the R_{sw} of the NPG is added to that of the inverter. The switching trajectories of an n-FET in an inverter and an NPG cover different sections of the I - V characteristic of a MOSFET, as shown in **Figure 5**. If the NPG width is small compared with those of the MOSFETs in the inverter, the delay is dominated by the R_{sw} of the NPG. On the other hand, if the reverse is true, the increase in switching capacitance from the overlap and diffusion regions of the NPG becomes significant. Similarly, information on the source-drain resistance is obtained from NAND and NOR circuit configurations [5].

We have implemented more than 120 ROs in the kerf covering a variety of logic gates such as inverters, NANDs, NORs, n-passgate and p-passgate circuits, and for extracting MOSFET parameters and parasitic components. A subset of these ROs is measured at the first level of metal (M1) for early learning and process tuning. In this case, each RO has an independent V_{dd} , and its output is fed to a circuit like that shown in Figure 3(b). Both IDDA and IDDQ of each RO can be measured directly. With three or more levels of metal, a more compact version of the test structure uses a five-bit decoder to select any one of 32 ROs which share a common output bus. The combined IDDQ of all ROs sharing a common V_{dd} is subtracted from the IDDA of the selected RO to obtain C_s using Equation (2).

All physical layouts are representative of product designs or have intentional variations to monitor the

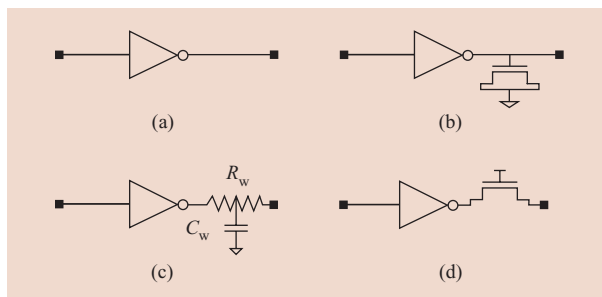


Figure 4

Ring oscillator stages: (a) reference inverter; (b) inverter with an n-FET gate load; (c) inverter with a wire load; (d) inverter with an n-FET passgate load.

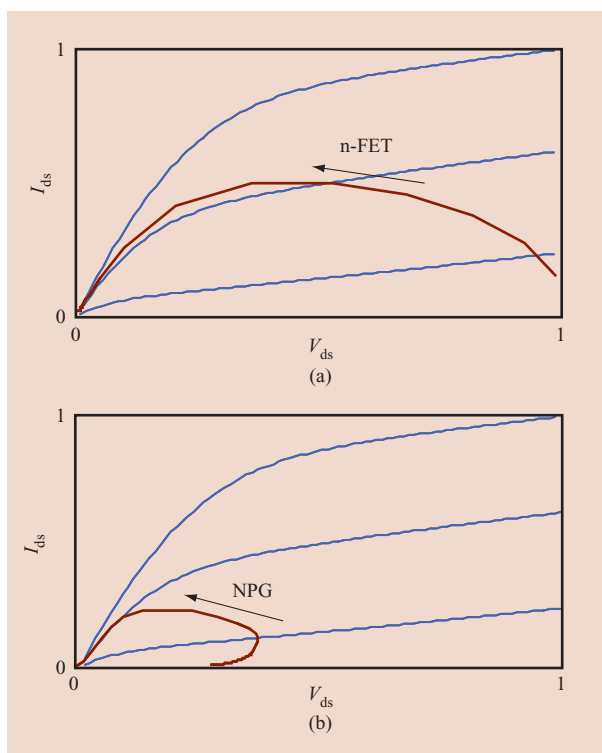


Figure 5

Trajectories in drain-to-source voltage (V_{ds})/drain-to-source current (I_{ds}) space of (a) an n-FET in an inverter and (b) an n-FET passgate during switching, superimposed on the n-FET dc I - V plots. Reproduced from [4], with permission; ©2005 IEEE.

sensitivity of performance and yield to physical layout. Data analysis is simplified by tracking all circuit types with a canonical inverter RO. Data is normalized to the specifications derived from circuit simulations of the ROs over a range of channel lengths, L_p , threshold voltages,

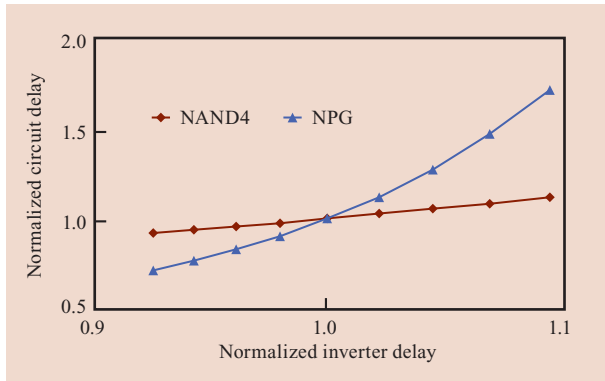


Figure 6

Simulated normalized change in delay for a NAND4 and an NPG circuit, with respect to normalized inverter delay for different n-FET V_t values.

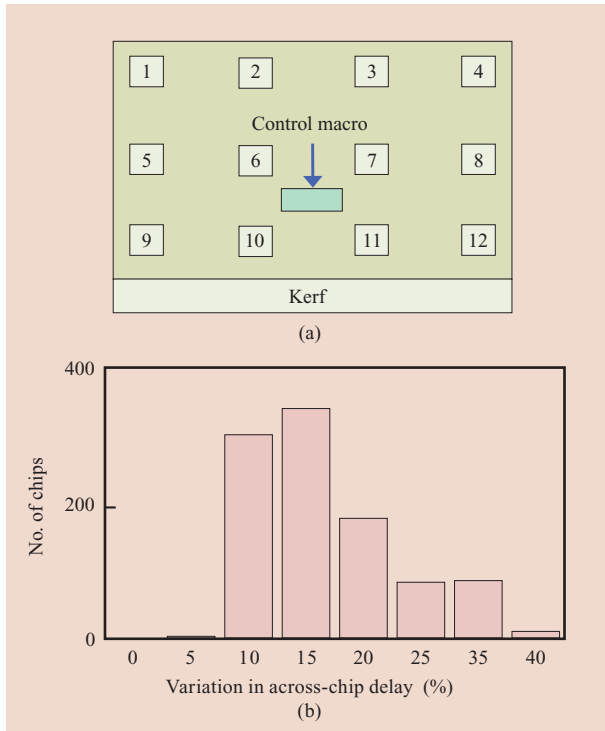


Figure 7

(a) Concept for physical layout of a product chip and the kerf, showing the location of 12 identical ROs distributed across the chip and a central control macro; (b) histogram showing range of delays of spatially distributed ROs for chips in a lot.

parasitic extraction. An example of delay tracking of two circuits in response to variations in V_t of the n-FETs is shown in **Figure 6**. One of the circuits is a four-input NAND, NAND4, with its top input switching; the second one is an n-passgate circuit. The delay is normalized to the delay with nominal parameter values and compared to a normalized inverter delay with the same variations. As expected, the NPG has a larger circuit delay variation range than the NAND4.

A selected set of these ROs is placed on the product in an arrangement shown in **Figure 7(a)**. Multiple copies of a standard fan-out-of-3 inverter are distributed across a product to track systematic delay variations. A single macro, placed near the center of the product or in close physical proximity to a critical area of the product, comprises 19 ROs of diverse circuit stages. It also contains the control circuitry, including a five-bit decoder, output OR, and frequency divider for all ROs on the chip. These ROs share I/Os with other product test functions such as built-in self-tests [7]. The RO frequencies are measured at wafer final test (WFT) as well as in packaged products. In contrast to the in-line tests, in which only a limited number of chips in a lot are measured, the data at WFT is collected on all chips on each wafer in a lot. Mapping these RO delays across a wafer, across the reticle, and across the product chip is extremely useful in getting to the root cause of delay spread, such as variations in optical masks, lithographic exposure, etch, rapid thermal annealing, and other process steps. Across-chip delay variations of chips within a single lot in the manufacturing line may vary considerably, as illustrated in **Figure 7(b)**. This across-product delay variation has been correlated with photon emission microscopic images of the full product chip [8] and with the variations in maximum operating frequency of different functional blocks. Identical ROs in the kerf and the product map the delay variation across the full reticle to ensure that the process tuning based on the in-line structures is equally applicable to that of the product. While the V_{dd} of the ROs in the kerf is isolated from the support circuitry to enable the measurement of RO capacitance, the ROs on the product itself are tied to the common V_{dd} , which precludes measurement of RO capacitance. Information on the MOSFET parameters and on resistance and capacitance variations is obtained by tracking the delays with respect to a canonical inverter [5]. The V_{dd} and temperature dependence of ROs with different circuit types are shown respectively in **Figures 8(a)** and **8(b)**. These, when compared with the V_{dd} and temperature dependence of the operating frequency of the product, provide insight into the frequency-limiting paths. Such techniques are very useful for identifying the sources of mismatches between model and hardware, and of variability with a resolution of about 3%.

V_t , and other MOSFET parameter variables [5]. The simulations are carried out for the entire RO using full

Ring oscillator small-signal gate capacitance test structure

In addition to the standard set of ROs described above, we have demonstrated a number of other RO-based test structures that address specific characterization challenges. An especially useful enhancement is the addition of an independent dc bias lead to each RO stage. Examples of this are shown in **Figures 9(a)** and **9(b)**, in which one of the MOSFET gates is connected to an adjustable potential instead of V_{dd} or GND as in comparable stages shown in Figures 4(b) and 4(d), respectively. One important application that makes use of a circuit such as that shown in Figure 9(a) is the measurement of small-signal capacitance–voltage (C – V) characteristics of thin-gate-dielectric capacitors.

As gate oxide thickness has been reduced to 2 nm and below, high parallel conductance associated with gate oxide tunneling has forced conventional gate capacitance measurements out of the purview of the in-line test environment. In particular, small-signal C – V characterization of standard-thickness gate oxides, which through the 180-nm node was routinely done as an in-line test, is now typically done as a 20-MHz bench test. An RO with the stage configuration shown in Figure 9(a), in which the source and drain of an n-FET are connected to the output of an inverter while the gate is connected to the independent voltage bias lead (V_{cg}), can be used to perform an in-line small-signal C – V analysis. The inverter power supply, V_s , is lowered to <0.5 V, and the delay and C_s of the RO are measured at different values of V_{cg} corresponding to bias voltages of $(V_{cg} - V_s/2)$. Here, V_s serves as a small-signal excursion on V_{cg} , and the entire C – V characteristic of the n-FET can be mapped out. The output driver is operated at 1.0 V to maintain the integrity of the signal to the external frequency counter, and voltage-level shifters are added to make the transition from a low to a high voltage. The inverter and other parasitic capacitances are eliminated, as previously described, by using a reference inverter RO of similar design and measured at the same value of V_s . This technique has the drawback that the “small-signal” voltage cannot be reduced below 150 mV in amplitude, but it has the tremendous advantage of rapid in-line determination of MOSFET gate capacitance in the presence of high gate conductance using only dc I/Os.

In **Figures 10(a)** and **10(b)**, simulated C – V plots of n-FETs with different values of V_t and L_p are shown for this experimental configuration in PD–SOI technology. As expected, the gate capacitance, C_g , in the negative bias depletion region, C_d , and in the large positive bias inversion region, C_i , are independent of V_t . The C_i and C_d values are measured in the flat regions of the C – V plot, so that the effect of the non-zero amplitude of the small-signal voltage is negligible. Here C_i gives a measure of

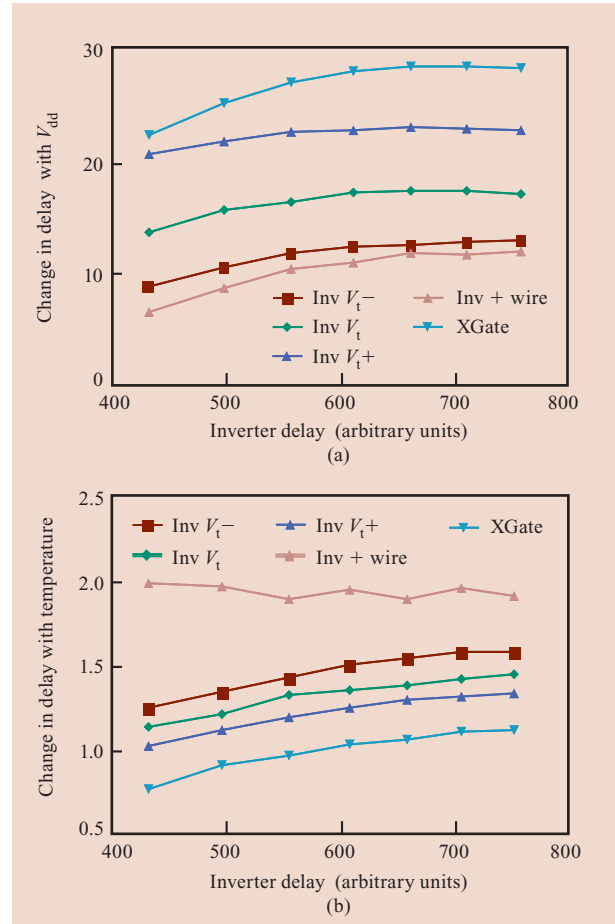


Figure 8

Change in delay with (a) V_{dd} and (b) temperature for different circuit types as a function of inverter delay in arbitrary units. The circuit types are inverters with different V_t s ($V_t^+ > V_t > V_t^-$), inverter with a wire load, and a transmission gate (XGate).

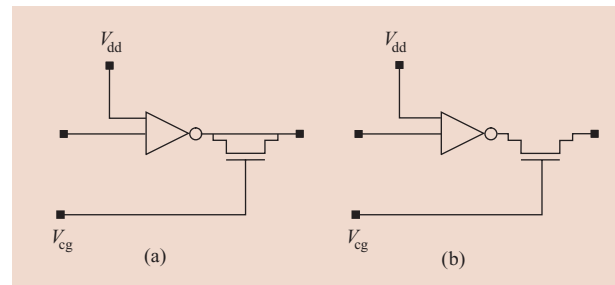


Figure 9

Circuits for RO stages with an independent voltage bias line, V_{cg} , connected to the gate of (a) an n-FET gate load; (b) an n-FET passgate load.

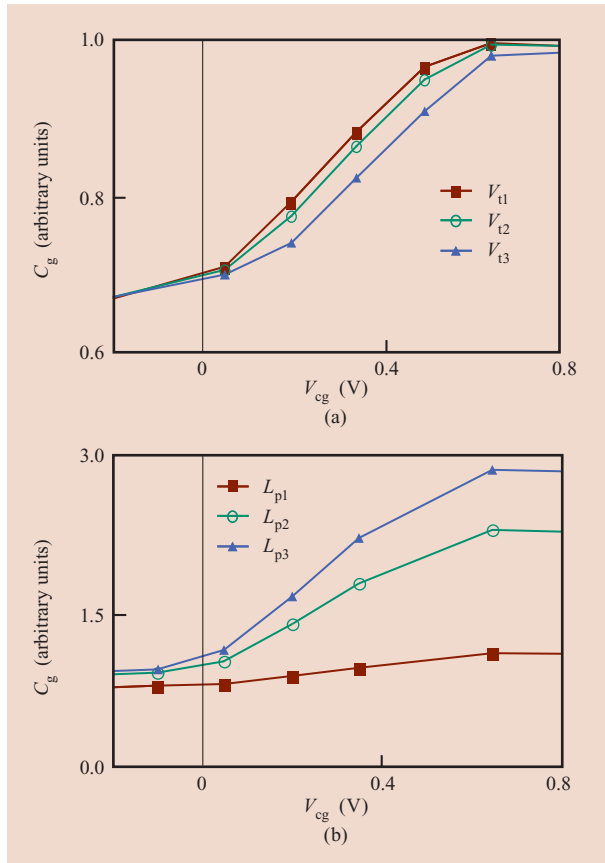


Figure 10

Simulated gate capacitance, C_g , vs. V_{cg} for (a) MOSFETs having different V_{ts} ($V_{t1} < V_{t2} < V_{t3}$), and (b) MOSFETs having different L_p s ($L_{p1} < L_{p2} < L_{p3}$).

polysilicon gate length. The influence of V_t is apparent in the transition region, as for a lower- V_t n-FET the channel is turned on at a lower V_{cg} . The midpoint on the C - V curve can be used to determine the value of the dynamic V_t . The C_i increases with an increase in L_p , as shown in Figure 10(b), while C_d , which comprises fringe, overlap, and parasitic capacitances, is nearly unchanged. The C_i at two different L_p values gives a measure of oxide thickness, T_{inv} , as follows:

$$T_{inv} = \frac{\epsilon\epsilon_0(L_{p2} - L_{p1})}{C_{i2} - C_{i1}}, \quad (4)$$

where ϵ is the dielectric constant of the oxide, ϵ_0 is the permittivity constant of free space, and C_{i1} and C_{i2} are the inversion capacitances per unit width at polysilicon lengths L_{p1} and L_{p2} , respectively. The effective capacitive channel length, L_{pe} , is estimated from $(C_i - C_d)$, since this is the additional capacitance due to the formation of the

inversion layer. The same design is adapted for p-FET characterization, and C_i is measured with negative values of V_{cg} . This unique RO design allows us to determine the delay, $IDDQ$, L_p , L_{pe} , T_{inv} , and V_t of a circuit and its associated MOSFETs in a self-consistent manner, providing a powerful new addition to the overall characterization arsenal.

This test structure has been implemented in both 90-nm- and 65-nm-technology nodes and is testable at the first level of metal. Using a modified version of the standard M1 testable RO design, ten ROs are accommodated per macro. One macro features both n-FETs and p-FETs with low, standard, and high V_{ts} , along with the reference and calibration inverters. The second macro features three different values of L_p for both n-FETs and p-FETs, again along with the reference and calibration inverters.

Ring oscillator beat-frequency test structure

One important source of variability in advanced CMOS technologies involves the change in performance over time due to various stress mechanisms, such as hot-electron effects in both n-FETs and p-FETs and negative bias temperature instability (NBTI) in p-FETs [9]. Developing an understanding of the dependence of such degradation on process details for each new technology is an important and difficult task involving accelerated stress testing of large populations of devices. It is of value to have in-line test structures that give an indication of the stress-induced performance degradation after only a few seconds of stress. A significant problem in doing this is that the signal is small and subject to error, especially in the case of NBTI at ambient temperature. We have developed a test structure in which the difference frequency between two nominally identical ROs in close proximity, and sharing a common power supply, is digitally calculated *in situ*. In practice, the beat frequency is measured, one of the ROs is briefly stressed, and then the beat frequency is measured again. The change in beat frequency, which is now a differential quantity, is a precise measure of the change of performance with stress and is relatively insensitive to factors such as power-supply variation and external noise. In addition, by configuring the RO stages in an appropriate fashion, it is possible to have a specific mechanism dominate the response of the circuit to stressing.

In this test structure, a beat-frequency generator, which comprises two latches together with combinational logic, delivers a signal at half the frequency difference between the two ROs, for frequency differences of up to about 25%. The individual frequency of each RO in a pair, as well as their beat frequency, is measured. There are two different implementations of this test structure. In the first

case, the power-distribution system for each RO connects to an independent I/O pad, but these pads are driven by a single common source for beat-frequency measurements. With this independent pad configuration, voltage stress can be applied to one RO by raising its V_{dd} while the other RO remains at nominal V_{dd} or lower. Degradation in circuit performance may result from both hot-electron effect and NBTI. Also, differential changes in contact resistance at the I/O contact pads may introduce errors in the measurements. Alternatively, the power-distribution systems of both ROs may be connected to a common I/O pad. Here the voltage stressing is accomplished by applying a voltage bias to the gates of MOSFETs in a configuration similar to that shown in Figure 9(b) while keeping the inverter power supply at ground potential. With the drain-to-source voltage, V_{ds} , of the NPG $\ll V_{dd}$, it experiences only NBTI stress, while the MOSFETs in the inverter are not stressed at all. The average change in V_t of the NPGs of the stressed ROs can be directly measured. It is equal in magnitude to the change in V_{cg} necessary to keep the beat frequency after stressing the same as it is prior to the stress.

The standard beat-frequency macro implementation has up to eight pairs of ROs with a common-frequency divider and output driver. A number of circuit types are represented, including inverters, NANDs, NORs, and n-passgate and p-passgate circuits. The numbers of stages in the ROs of a pair are 104 and 100, with the 104-stage RO being stressed. This small inequality ensures that the beat frequency always increases after stress. The voltage stress effects are easily measured for room-temperature stress times as short as ten seconds, as shown in **Figure 11(a)** for p-passgate circuits. In addition, the distribution in the beat frequency of similar pairs of ROs provides information on the random variations in MOSFET parameters. As shown in **Figure 11(b)**, an RO pair using a p-passgate circuit shows a wider spread in beat frequency than an RO pair of three-input NAND (NAND3) gates. This increased spread is attributed in part to the high sensitivity of the p-FET pass-transistor to random V_t variations.

Test structures for measuring local random variability

With continued scaling of CMOS technology, local random variability in parameters such as V_t and the resistance of metal-to-MOSFET contacts (R_{ca}) is reducing operating margins and compromising power/performance tradeoffs [10, 11]. Existing comprehensive test sites for measuring statistical variations in MOSFET parameters typically require several hours of test time in an off-line test environment. It is thus of value to develop rapidly testable compact test structures to measure

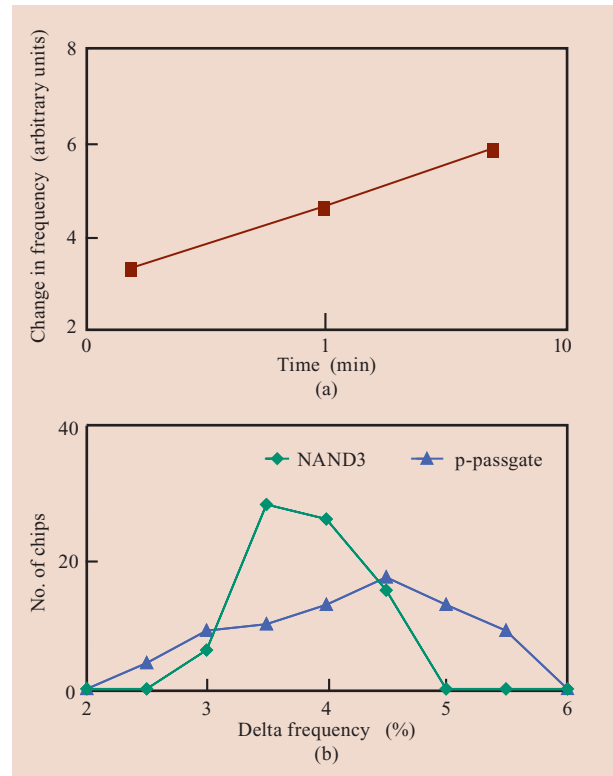


Figure 11

(a) Change in beat frequency with stress time, where only one RO is stressed at 1.8 V, 25°C; (b) distributions of measured beat frequency for a pair of ROs with a built-in offset of 4%. The NAND3 has a smaller spread than an inverter driving a p-FET passgate.

aspects of local random variation in-line on a routine basis, for providing rapid feedback for technology development and manufacturing.

We have developed new test structures that address this characterization need. **Figure 12** shows the basic concepts of these designs. The underlying idea is to have an internally driven addressing scheme, an array of circuits or circuit components, the measured property of which is sensitive to some physical parameter of interest, and a readout scheme that utilizes the statistical measurement capability of instrumentation already present in existing in-line test equipment. For the case shown in Figure 12(a), a clock RO drives a frequency divider, followed by a counter with p stages. The outputs of the counter drive the inputs of a decoder, which in turn provides sequential activation signals to the 2^p elements of an array under test. The array outputs are multiplexed together onto a common output bus. The final signal follows the output of the sequentially activated array elements. If the array elements are ROs, the common

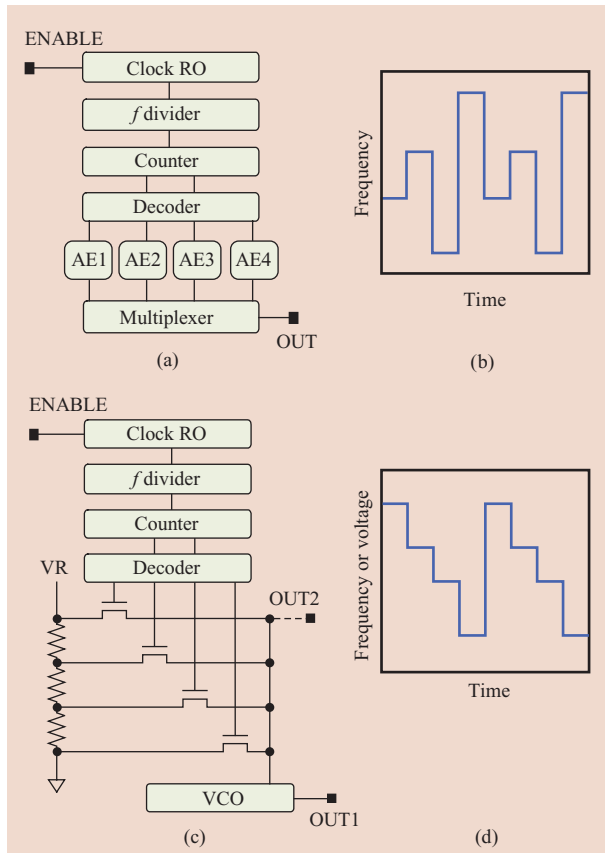


Figure 12

(a) Circuit scheme for measuring parameter distributions of array elements AE1 to AE4; (b) frequency-modulated output waveform; (c) circuit scheme for measuring resistance distributions; (d) output waveform.

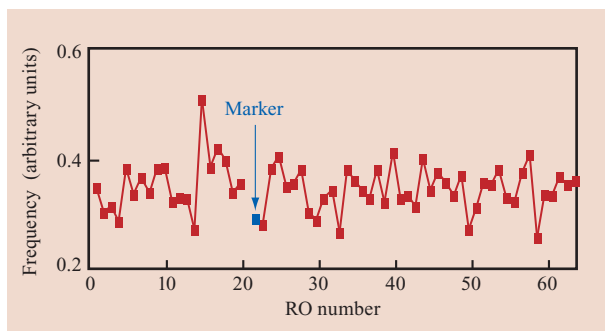


Figure 13

Measured frequency of individual ROs in a 4×16 array. One position in the array is left blank, serving as a marker to identify the physical location of the ROs.

output is a frequency-modulated signal, as indicated in Figure 12(b). This frequency-modulated signal is in turn measured and analyzed by the existing in-line frequency counter, either alone or in combination with local data processing. In other applications, the output may be a current or voltage signal that is measured with a standard in-line current meter or voltmeter having statistical measurement capability.

A second arrangement is shown in Figure 12(c), in which the array elements are, for example, metal via contact resistances. Here the decoder sequentially addresses points down the resistive ladder voltage biased at V_r . The voltages at these points drive the input of a voltage-controlled oscillator (VCO), the output of which (OUT1) is a frequency staircase as shown in Figure 12(d), where the step heights correspond to individual via resistances. Alternatively, the VCO could be replaced with a unity gain buffer for a direct voltage readout from OUT2.

With an RO array approach, the design of the RO stage is made sensitive to one selected parameter so that the statistics of the frequency-modulated output relate back directly to the statistics of the physical property being investigated. An example is the RO stage shown in Figure 4(d), which is sensitive to the V_t of the NPG. For measuring variations in R_{ca} , the RO stage comprises a wide logic gate with a chain of 20 metal-to-MOSFET contacts in series with its output, such that the frequency of the RO is sensitive to anomalous changes in R_{ca} . For a compact design, the ROs typically have five stages, including a NAND2 gate for enabling the RO. Eight such arrays, each with 64 or 128 ROs, are included in a single macro.

The measured frequencies of each of the members of an array of 63 n-passgate ROs are shown in Figure 13. One RO location of this 4×16 configuration is intentionally unused and serves as a physical marker. The variability in RO frequency arises from variations in the RO stages. The variations in the five stages in each RO are assumed to be statistically independent of one another. It follows that the measured variability represents an average over five stages so that the variation-induced frequency change, δf , is reduced by a factor of $\sqrt{5}$ from what it would be if all of the stages in each RO varied together. In addition, the statistical integrity of the results improves as the number of members in the array increases. The V_{dd} dependence of the stage delay is dominated by the NPG. Since for the NPG a change in V_{dd} (δV_{dd}) of the gate voltage is equivalent to a negative change in V_t of the same magnitude, it follows that $\delta f / \delta V_{dd} \approx -\delta f / \delta V_t$. Alternatively, with the gates of the NPG connected to an independent voltage source, as shown in Figure 9(b), $\delta f / \delta V_{cg} = -\delta f / \delta V_t$. Either $\delta f / \delta V_{dd}$ or $\delta f / \delta V_{cg}$ can be

directly measured and used to extract the V_t statistics from the frequency data.

Test structure for electrically programmable fuse diagnostics

In recent CMOS technologies, an electrically programmable fuse, eFuse, has replaced laser-blown fuses for memory redundancy and permanent storage of custom chip information [12]. The eFuse comprises a standard polysilicon gate stack having a small resistance in its pre-blown state. The fuse is blown with the application of a high voltage, typically 3.5 V, via electromigration effects. To monitor the integrity of the eFuse design and process early in the manufacturing cycle, it is important to carry out the fuse blow in a manner representative of the product environment, along with measuring pre-blow and post-blow fuse resistance. In a product, the fuse blow process involves a pulse with a rise time of 1 ns or less and a duration of a few hundred microseconds. Pulse generators to produce such pulses can, for a significant capital investment, be added to the in-line testers, but this still leaves concerns over the shape of the waveform delivered through the dc I/O input line.

As a possible solution to this characterization challenge, we have designed a test structure for in-line characterization of eFuse using parametric testers with only dc I/Os [13]. A key component of this test structure is the pulse generator. The pulse width required for this application is five orders of magnitude greater than what can be practically generated with the scheme shown in Figure 3(d). A simplified schematic diagram of the pulse generator circuit developed for this application is shown in **Figure 14**. A ring oscillator with 241 stages is enabled by setting the input $ENABLE = "1"$ and serves as an on-chip clock. A "dc" Launch signal creates a sharp rising edge for the pulse and initiates a resettable counter. At a selectable time after the generation of the first pulse edge, a signal is sent to create the falling edge of the pulse. A latch, LatchA, is used for creating the rising edge of the pulse, as shown in Figure 3(c). With a "1" preloaded into the data port of LatchA (via input Arm) and with its output at "0," all is quiet until the Launch input to the LatchA clock is of sufficient magnitude to allow the loaded "1" to pass to the output. The only requirement for the Launch input signal is that it must undergo a transition from "0" to "1" (the details of the waveform, including the duration of the transition, are unimportant). The output from LatchA is a very sharp edge that occurs at some point during the rise of the Launch signal, and subsequent events are self-timed with respect to this sharp edge. The LatchA output forms the leading edge of the pulse for the eFuse blow at the OUT terminal, preloads a "1" into the data port of LatchB, and also turns off the reset signal (r) to the resettable counter.

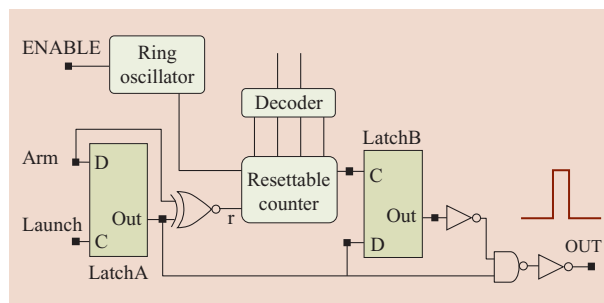


Figure 14

Circuit scheme for generating a sharp but wide pulse for eFuse characterization. Here C and D are the clock and data inputs of LatchA and LatchB.

Table 1 Output pulse widths for different decoder settings.

$a1$	$a2$	Pulse width
1	1	0.7 μ s
1	0	11 μ s
0	1	200 μ s
0	0	3.2 ms

Table 2 Sequential inputs for initializing the pulse generator circuit and generating a pulse.

Time step	ENABLE	Arm	Launch
1	1	0	0
2	1	0	1
3	1	0	0
4	1	1	0
5	1	1	1

The counter counts up to a specific time determined by the decoder inputs $a1$ and $a2$, and then sends a signal to the clock input of LatchB, which is waiting with the preloaded "1." Next, the output from LatchB is inverted and combined with the original LatchB output to form the falling edge of the pulse, which appears at the OUT terminal and is applied across the fuse. The resistance of the fuse is measured before and after the application of the pulse to quantify the performance of the eFuse structure. **Table 1** gives an example of output pulse widths for various decoder inputs.

Table 2 shows the set of sequential inputs necessary to initialize the circuit and generate the pulse. Each step is a few ms in length. Steps 1 through 3 reset the latches,

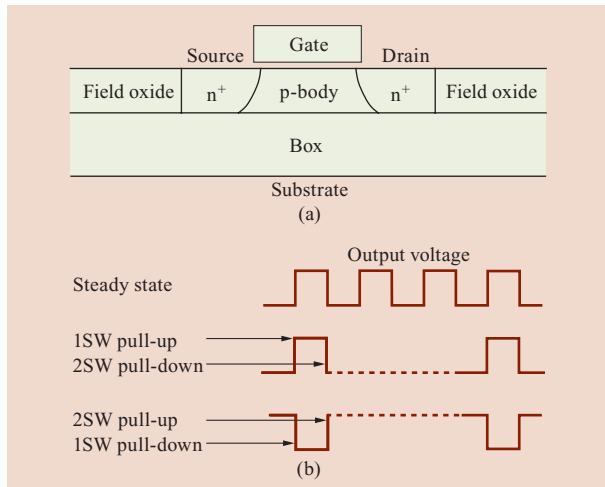


Figure 15

(a) Physical cross section of an n-FET in PD-SOI technology; (b) waveforms showing steady state (SS), 1SW, and 2SW transitions for PD-SOI delay history measurements. Part (b) reproduced from [19], with permission; ©2004 AIP.

ensuring that there is a “0” at the latch outputs. Note that during step 2 the counter is counting and the clock input to LatchB is oscillating. Since the data input to LatchB is a “0” at this time, the LatchB output is set to “0.” In step 4, a “1” is preloaded into the LatchA data port and then step 5 creates the pulse as previously described. While generating a pulse for eFuse characterization is one possible application for this pulse generator, there are many other applications that can benefit from such variable-width pulse-generation capability.

Test structures for PD-SOI technology

The cross-sectional diagram of a PD-SOI n-FET is shown in **Figure 15(a)**. PD-SOI technology presents unique characterization challenges. The buried oxide introduces a thermal barrier between the active device area and the underlying substrate, enhancing temperature excursions, as the device is turned on. In addition, the floating-body potential is determined by a combination of long-time-constant processes including diode leakage and gate-to-body tunneling, along with very rapid processes involving various components of capacitive coupling. This floating-body potential in turn influences the V_t of the MOSFET and consequently its performance [14]. Because of floating-body effects and self-heating, measurement results depend critically on the recent history of operation of the device under test. For example, dc $I-V$ characteristics as commonly acquired and tracked in-line are intrinsically different from the

corresponding characteristics under high-speed switching conditions [15, 16].

Output waveforms from an individual logic gate that correspond to switching scenarios commonly used to characterize circuit delay history are shown in **Figure 15(b)**. The first waveform in this figure is used for determining the steady-state (SS) delays with a square-wave input. For the second waveform, a pull-up transition occurs after a long period of rest, typically of the order of a few ms or more, and the pull-down transition occurs within a few ns or less following the first transition. These transitions are depicted as first switch (1SW) and second switch (2SW), respectively. Here the pull-up transition is defined as a “0” to “1” transition at the output node of a CMOS gate and is dominated by the p-FET current drive capability. Similarly, the pull-down transition is defined as a “1” to “0” transition at the output node of a CMOS gate and is dominated by the n-FET current drive capability. In the third waveform, the 1SW and 2SW transitions correspond to pull-down and pull-up instead. The fractional difference between 1SW and 2SW delays is the 1SW-2SW history, with the 1SW-SS and 2SW-SS histories similarly defined. In current PD-SOI technologies, 2SW transitions can be 10% or more faster than 1SW transitions [14].

The RO-based test structures described previously operate only under SS conditions. It is thus of considerable value to develop test structures that allow direct comparison of behavior under a variety of bias conditions and switching sequences. In the following sections, a number of test structures are described that address these characterization challenges.

In-line measurement of pulse $I-V$ characteristics

MOSFET dc $I-V$ characteristics are used extensively in the characterization of both bulk silicon and PD-SOI technologies. As stated above, the $I-V$ characteristics relevant to high-frequency operation in PD-SOI differ from dc $I-V$ behavior as a result of both self-heating and floating-body effects. High-frequency pulse-based $I-V$ characterization of MOSFETs in PD-SOI technology has traditionally been carried out as an off-line bench test. Because of the time-intensive effort required, these measurements are done infrequently and have not played a significant role in the technology optimization process. We now describe a test structure for in-line measurement of pulse $I-V$ characteristics using only dc I/Os, providing the ability to directly compare these characteristics with standard dc characteristics [17].

The basic principle of this new measurement technique is illustrated in **Figure 16** for the case of n-FETs as the devices under test. A continuous sequence of non-overlapping pulses is generated sequentially from an RO and applied to the gates of ten nominally identical

n-FETs under test. The n-FETs are connected in parallel, and the measured current is averaged over all n-FETs, each with a 10% duty cycle. The RO consists of 1,000 stages and is divided into ten equal segments, each 100 stages in length. The output from the XNOR across each segment is a train of complementary pulses of width $T/10$ and period T , where $T = 1,000d$ and d is the delay per stage of the RO. The RO and XNORs share a common V_{dd} and a common GND. The pulse trains from the ten segments are non-overlapping, of magnitude V_{dd} , and configured such that at any given time one and only one complementary pulse is present.

The output from each XNOR drives an inverter with an independent power supply of amplitude V_g . The output of each such inverter in turn applies V_g to the gate of an n-FET under test with its drain held at V_{ds} and its source at GND, as shown in Figure 16(a). Each n-FET is on and drawing drain current 10% of the time and is off the remaining 90% of the time. The current drawn from the V_{ds} supply is nominally constant and equal to that of a single n-FET under test. To the external current meter, this I_{ds} (ac) appears as a constant dc current which can be accurately measured. Since each n-FET is only on and carrying current 10% of the time, the self-heating, with a time constant of ~ 100 ns, is 10% of that experienced by a similar device under conventional dc test conditions, and rendered insignificant. Furthermore, the floating-body potential of the MOSFETs under test is very nearly that of similar devices in a high-duty-factor steady-state operation under representative use conditions. With the V_{dd} of the RO set to GND, the outputs of the XNORs are at GND potential, and the outputs of all of the intermediate inverters applied to the gates of all of the MOSFETs under test are at V_g . The I_{ds} then measured as the dc current from the V_{ds} supply is the standard I_{ds} (dc) for the ten n-FETs in parallel. This circuit thus allows one to directly measure and compare, for the same devices, the standard I_{ds} (dc) with the corresponding pulse value, I_{ds} (ac). For characterization of p-FETs, the voltage bias polarities are reversed.

This circuit has been implemented as an in-line macro with two 1,000-stage ROs, one driving two groups of ten n-FETs and the other driving two groups of ten p-FETs. Care is taken to ensure precise temporal alignment of pulses, and low- V_t MOSFETs are incorporated in the inverters preceding the FETs under test to extend the operating range to $0.5 < V_g/V_{dd} < 1.3$. The resulting overall error in the measured I_{ds} (ac) due to misalignment and skew is less than 1%. The ring oscillator frequency can also be measured and correlated with MOSFET characteristics. Data from experimental hardware illustrates the expected suppression of the I_{ds} (dc) values at high V_{ds} and V_g [17].

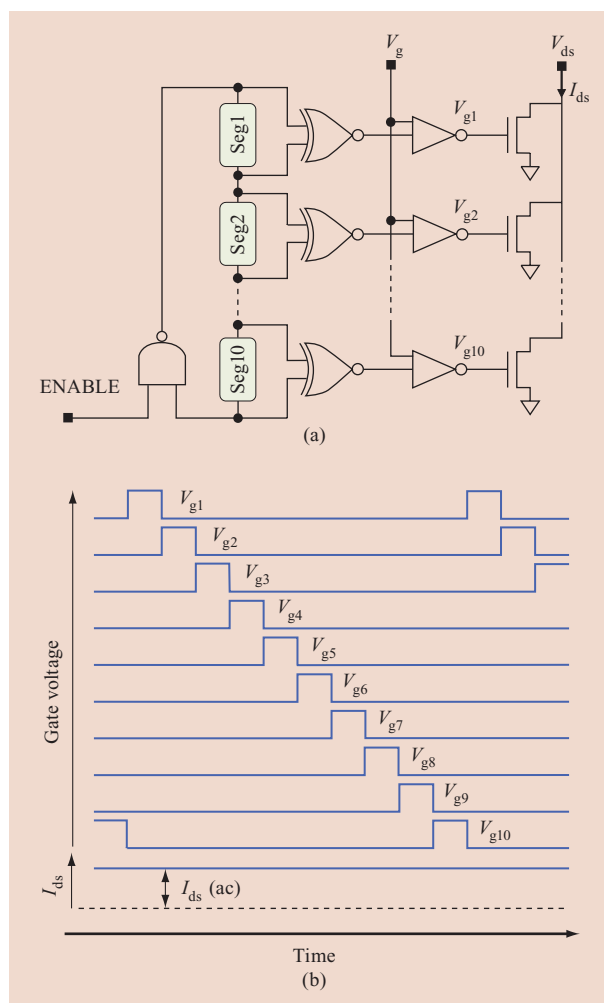


Figure 16

(a) Circuit scheme for in-line pulse I - V measurements; (b) non-overlapping pulse inputs to the gates of the ten n-FETs under test and the associated constant I_{ds} (ac).

In-line measurement of PD-SOI history

The measurement of floating-body-induced switching history effects in PD-SOI has traditionally required the use of off-line high-frequency bench test equipment. Since history effects can modulate logic gate delay by 10% or more, as well as contributing to SRAM cell instability, it is valuable to have an in-line measurement of switching history, preferably one that can be done at the first metal level for rapid process optimization. Such an in-line test structure for determination of average switching history in delay chains using only dc I/Os has been implemented in both the 90-nm- and 65-nm-technology nodes.

This test structure utilizes the property that the width of a pulse launched at the beginning of a long delay chain of PD-SOI gates changes as it travels to the other end

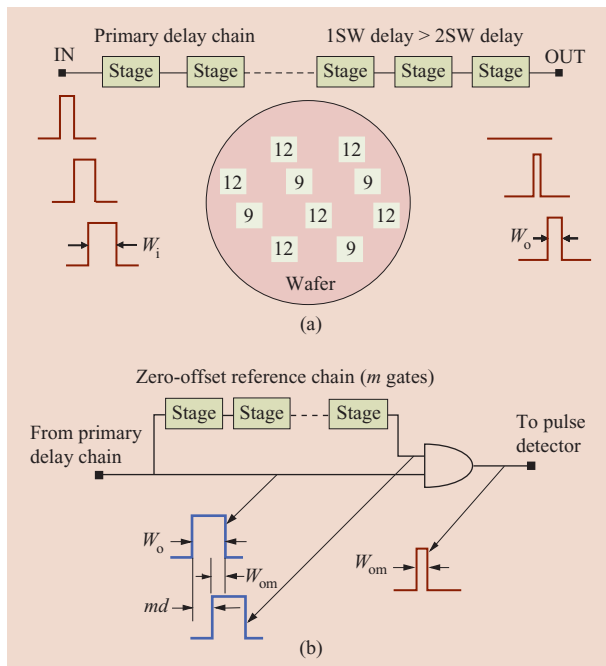


Figure 17

(a) Illustration of in-line delay history measurement concept; inset shows hardware data format indicating history values of 9% and 12% for the ten chip sites measured; (b) zero-offset circuit used with in-line history test structure to enable measurement of negative history.

[4, 18]. **Figure 17(a)** shows a long delay chain of nominally identical PD-SOI stages, where a stage could be an inverter or some more complex circuit. When a pulse of sufficient width W_i is launched, it emerges from the far end of the delay chain with a width of W_o . The history can be defined as $(W_i - W_o)/D_{ch}$, where D_{ch} is the time for the first edge of the pulse to travel the length of the delay chain. The delay per stage of this first edge is the 1SW delay, and the delay per stage of the second edge of the pulse is the 2SW delay, both averaged over pull-up and pull-down. The average (over pull-up and pull-down) 1SW–2SW history is >0 if $W_o < W_i$, which is the case for all IBM PD-SOI technologies at their nominal operating voltages. As a realistic example, if $W_i = 2$ ns, $W_o = 1$ ns, and $D_{ch} = 10$ ns, the 1SW–2SW history is 10%.

As suggested in **Figure 17(a)**, an alternative way to determine the history is to launch a number of pulses with different W_i values. For some critical initial pulse width W_{crit} , the pulse will be annihilated just as it reaches the far end of the delay chain. The 1SW–2SW history can then be expressed as W_{crit}/D_{ch} . This scheme, together with the circuit blocks shown in **Figure 3**, is used as the basis for an in-line history test structure. The circuit shown in **Figure 3(c)** is used to generate a single sharp

edge, in response to a slow-rising clock input signal, which is launched down the primary delay chain and also into a shorter parallel reference chain made up of the same stages as the primary delay chain. After traversing a pre-selected number of stages in the reference delay chain, the signal is fed back to the beginning of the primary delay chain, returning the input of the primary delay chain to “0” and forming the 2SW edge. Another latch circuit of the type shown in **Figure 3(c)** is positioned at the far end of the primary delay chain, with a preloaded “1” in its data port and its clock input connected to the chain output. If the launched pulse arrives at the end of the chain, the clock is momentarily high, and the “1” is passed to the latch output, where it can be viewed at a later time. A three-bit decoder is used in conjunction with the reference delay chain to create eight different values of W_i , which are sequentially launched after successive latch read and reset cycles. In practice, the primary delay chain has 1,200 stages. The reference delay chain has eight segments, each with 36 stages, corresponding to 3% of the total delay of the primary chain. The W_i/D_{ch} values thus range from 3% to 24% in steps of 3%, and the latch output is a set of “0” values followed by a set of “1” values, where the transition from “0” to “1” defines the history to within 3%. For example, “00011111” implies that the history is $>9\%$ and $<12\%$.

Another useful feature that can be added to the in-line history test structure is a zero-offset circuit, shown in **Figure 17(b)**, which can be used to enable measurement of negative history values. This feature is relevant at low values of V_{dd} , where the 1SW–2SW history can be negative. The circuit, which is similar to the pulse generator circuit shown in **Figure 3(d)**, is inserted between the end of the primary delay chain and the output latch. It contains a zero-offset reference chain, an even number m standard stages in length, and shortens the length of the output pulse from W_o to W_{om} , where $W_{om} = W_o - md$, and d in this case is the 1SW gate delay. Thus, for example, if $m = 108$, with a 1,200-stage primary delay chain, the circuit now measures history in the range of -6% to 15% in 3% increments.

This test structure is self-calibrating, since the delay of each reference chain setting scales with the delay of the primary delay chain. It is self-timed because the timing of the falling edge with respect to the rising edge of the input pulse is precisely determined by one of the internal feedback paths, the exact timing of the initial input pulse being unimportant. The input and output signals are low frequency, essentially dc. The raw output data, the format of which is shown as an insert in **Figure 17(a)**, is directly used as an input for process optimization (i.e., big is bad, small is good). Versions of the basic history test structure testable at the first metal level have been implemented for a variety of inverters with different values of V_t , different

device widths, and different device-width ratios. With four metal layers, circuits under test have been expanded to include NANDs, NORs, n-passgate and p-passgate circuits, and a variety of SRAM cell components.

Test structures with high-frequency I/Os

While in-line testable structures are very valuable for tracking technology and rapid diagnostics in the manufacturing line, bench tests still command an important position in any overall characterization strategy, particularly in early technology development and for model building. Motivated by a need for more comprehensive floating-body analysis, we have developed a scheme for measurement, with sub-picosecond precision, of circuit delays that are independently dominated by the pull-up (p-FET) and pull-down (n-FET) characteristics of a single- or multiple-input gate [19–21]. Any combination of inputs and number of switching events, arbitrarily configured with respect to timing and sequence, may precede the event to be measured. This scheme is used to characterize PD–SOI history effects in individual circuits, as well as a variety of other high-speed effects related to both SOI and bulk technologies.

A circuit diagram illustrating the basic concepts of this measurement scheme is shown in **Figure 18(a)**. In this example, two different experiments are multiplexed into a single circuit block. In the first experiment, the difference in delay between an unloaded and a loaded NAND3 is measured. In the second experiment, the difference in delay between a chain of 15 NAND3s and a chain of five NAND3s is measured. Five high-speed inputs A, B, C, D, and F are used both to select the experiment to be exercised and to implement a desired switching sequence. For example, with $B = C = F = 1$ and $D = 0$, the switching behavior of a top-switching NAND3 can be studied by toggling input A (V_{IN}) between “0” and “1.” Input S is used to select either the upper or lower path of the experiment under test. The pulse generator that initiates a switching event also triggers a sampling scope that captures the OUT signal. As shown in the waveform sketch in **Figure 18(a)**, as input S is toggled, the OUT waveform (V_{OUT}) shifts back and forth in time, with no change in shape. In the first experiment, the difference in time delay, δD , is approximately equal to the switching resistance, R_{sw} , times the known capacitive load CL. In the second, δD is equal to the delay of ten NAND3s well away from the ends of the chain. This shift is measured with sub-picosecond precision using a standard off-the-shelf sampling scope with 20-GHz bandwidth. A custom-designed wide-bandwidth probe card with six GND–Signal–GND probes for high-speed I/O (five inputs and one output) is used. Both of these experiments are intrinsically difference experiments configured such that

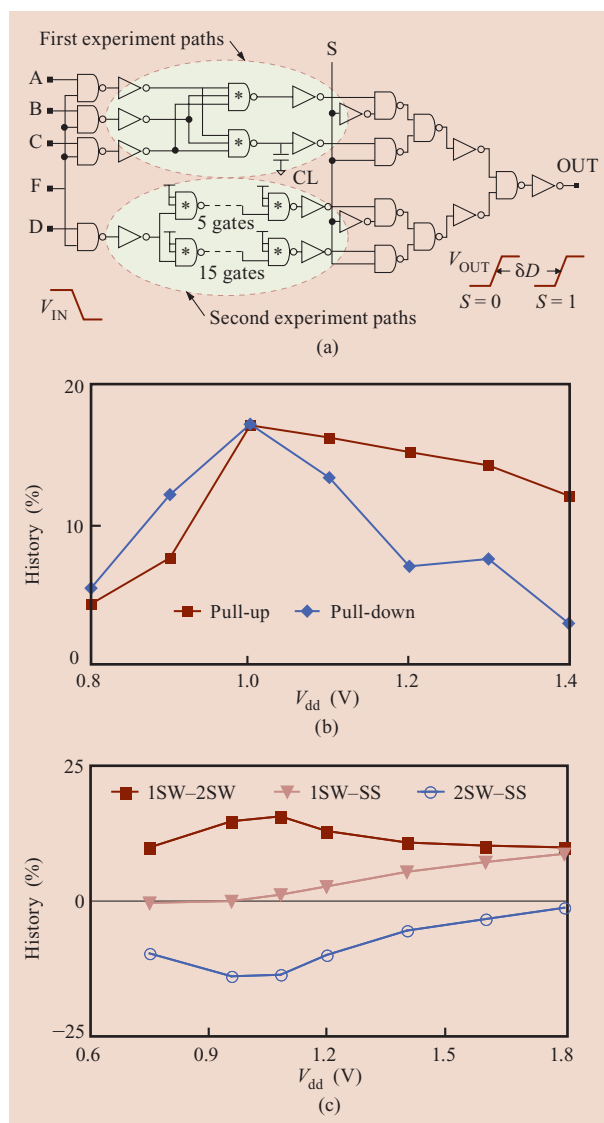


Figure 18

(a) Circuit scheme for sub-picosecond, time-resolved delay measurements, with symbols enclosing asterisks representing circuits under test. (b) Hardware data showing inverter 1SW–2SW history for pull-up and pull-down delays as a function of V_{dd} . (c) Inverter chain delay history for 1SW–2SW, 1SW–SS, and 2SW–SS transitions as a function of V_{dd} .

many sources of error are directly subtracted out. Just as a frequency shift can be accurately measured, here the time shift can be measured with high precision and stability.

With the present test structure, the pull-up and pull-down delays can be measured separately using the waveforms shown in **Figure 15(b)**. Beginning with a 100-MHz square wave, the pulse period is increased while keeping the pulse width constant, for both polarities of

input pulses. The 1SW and 2SW pull-up and pull-down delays are measured as a function of the pulse period, providing a measure of various time constants associated with the history. As the period becomes sufficiently large (typically $>10\text{--}100\ \mu\text{s}$), the delays become independent of period, with the fractional difference between 1SW and 2SW delays then defined as the 1SW–2SW history, the 1SW–SS and 2SW–SS histories being similarly obtained.

Figures 18(b) and 18(c) show data from experimental hardware measured with this scheme. In Figure 18(b) 1SW–2SW pull-up and pull-down histories for a loaded vs. unloaded inverter experiment in 90-nm PD–SOI technology are shown as a function of the power-supply voltage V_{dd} . In this case the pull-down and pull-up histories are similar for $V_{\text{dd}} < 1.0\ \text{V}$. Above 1.0 V, however, the pull-down history decreases much more rapidly as V_{dd} is increased than does the pull-up history. The data in Figure 18(c) is taken from a 15–5-inverter-chain experiment similar to the second experiment in Figure 18(a) in 130-nm PD–SOI technology. In this case, as with the in-line history experiment, it is the average of the pull-up and pull-down histories that is measured. 1SW–SS, 2SW–SS, and 1SW–2SW histories are all plotted as a function of V_{dd} , where the nominal operating $V_{\text{dd}} = 1.2\ \text{V}$. The 1SW–2SW history peaks at around 1.1 V. It is noteworthy that at low V_{dd} most of the history is associated with 2SW speedup compared with SS, while at high V_{dd} most of the history involves the slowdown of 1SW with respect to SS. At $V_{\text{dd}} = 1.2\ \text{V}$, SS delay (which is also what is measured with an RO) is about 3.5% longer than the average of 1SW and 2SW delays, while at $V_{\text{dd}} = 1.8\ \text{V}$ it is 3.5% less.

As test structures are migrated from one technology node to the next, steps are continuously taken to improve the overall efficiency of space utilization. In 65-nm technology, the bench test history macro dimensions are $165\ \mu\text{m} \times 2,500\ \mu\text{m}$, including the I/O pads. The macro template accommodates eight independent circuit blocks such as that shown in Figure 18(a). A distributed decoder is used to select the circuit block that will be active. A single analog input can be used in conjunction with current-starved inverters [7] to provide adjustable slews for the high-speed inputs or as an independent power supply for selected experiments. The eight independent circuit blocks share a common V_{dd} , while I/O circuits and buffers are on a separate power supply, V_{dd} (I/O), with a common GND. There are ten GND pads, with every other pad being a GND in the high-speed I/O region. On-chip decoupling capacitors are provided to ensure minimal power-supply droop during circuit-switching activity. This, together with the differential time domain measurement technique, provides robust results verified to be independent of the value of the V_{dd} (I/O). A large

number of experiments (as many as 32) with a variety of different circuit and device types can be accommodated within a single macro. While many of the designs focus on floating-body studies of standard static logic gates, exercising all members of the device menu, a number of experiments addressing SRAM floating-body effects are also included, such as experiments to measure minimum operating voltage as a function of write-to-read delay, and direct measurement of minimum wordline write pulse width. There are, in addition, a number of other experiments for studying phenomena ranging from latch metastability to signal crosstalk to dynamic adjacent and self heating.

Summary

A set of test structures has been developed for characterizing CMOS technology in a way that couples both to the underlying device physics and parametrics and to the product performance. The guiding themes in the design of these test structures are product-representative circuits, differential techniques, “at speed” functions, and dc I/Os. The structures are directly derived from circuits used in IBM products. A variety of differential design, measurement, and analysis techniques are used. While structures for standard MOSFET dc characterization are also included, the focus is on structures that operate at speeds representative of multi-GHz microprocessors. The emphasis is on designs that are fully operational with low-frequency inputs and outputs compatible with standard in-line parametric testers, although a subset of structures that leverage sub-ps time-resolved measurements is also included. Such product-representative ring oscillators and pulse-based structures have been placed on the kerfs of IBM PD–SOI chips starting with the 180-nm-technology nodes, and are used by IBM and its alliance partners to both characterize and evaluate the technology. Some of the test structures are also embedded in the microprocessors, and the complete set forms a basis for “at speed” performance, power, and variability characterization of the technology.

Acknowledgments

The authors gratefully acknowledge their collaboration with Carl J. Anderson, Stuart Bermon, K. K. Das, Anne Gattiker, Keith A. Jenkins, Dale J. Pearson, and Stas Polonsky on the design and analysis of the test structures, and Robert Havreluck and Steven Klepner for their assistance in physical design. We deeply appreciate the contributions of many other individuals in the IBM Research Division and the IBM Systems and Technology Group in various aspects of design, implementation, test, and analysis of these test structures.

Appendix A







Symbol	Definition
k	Divide-by factor
$2n$	Number of identical RO stages
d	Delay of a stage
f	Frequency of oscillation
IDD _A	Active current
IDD _Q	Quiescent current
C_s	Capacitance of an RO stage
R_{sw}	Switching resistance
L_p	Physical channel length
V_t	Threshold voltage
V_{cg}	Gate bias voltage
V_{ds}	Drain-to-source voltage
I_{ds}	Drain-to-source current
C_L	Load capacitance
R_{ca}	Metal-to-MOSFET contact resistance
	n-FET
	Inverter
	Two-input NAND
	Two-input AND
	Two-input NOR
	Two-input XNOR

Figure 19

Definitions of common symbols used in the paper.

References

- J. S. Panganiban, "A Ring Oscillator Based Variation Test Chip," M. Eng. Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, May 2002.
- B. E. Stine, E. Chang, D. S. Boning, and J. E. Chung, "Analysis and Decomposition of Spatial Variation in Integrated Circuit Processes and Devices," *IEEE Trans. Semicond. Manuf.* **10**, 24–41 (1997).
- A. Bassi, A. Vegetti, L. Croce, and A. Bogliolo, "Measuring the Effects of Process Variations on Circuit Performance by Means of Digitally-Controllable Ring Oscillators," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, 2003, pp. 214–217.
- M. Ketchen, M. Bhushan, and D. J. Pearson, "High Speed Test Structures for In-Line Process Monitoring and Model Calibration," *Proceedings of the IEEE International Conference on Microelectronic Test Structures*, 2005, pp. 33–38.
- M. Bhushan, A. Gattiker, M. Ketchen, and K. K. Das, "Ring Oscillators for CMOS Process Tuning and Variability Control," *IEEE Trans. Semicond. Manuf.* **19**, 10–18 (2006).
- Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, Cambridge, UK, 1998, Ch. 5.
- N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design*, Addison-Wesley Publishing Co., New York, 1992, p. 366, pp. 465–506.
- S. Polonsky, M. Bhushan, A. Gattiker, A. Weger, and P. Song, "Photon Emission Microscopy of Inter/Intra Chip Device Performance Variations," *Microelectron. Reliabil.* **45**, 1471–1475 (2005).
- D. K. Schroder and J. A. Babcock, "Negative Bias Temperature Instability: A Road to Cross in Deep Submicron CMOS Manufacturing," *J. Appl. Phys.* **94**, 1–18 (2003).
- A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulations and Intrinsic Parameter Fluctuations in Decanometer and Nanometer-Scale MOSFETS," *IEEE Trans. Electron Devices* **50**, 1838 (2003).
- B. P. Wong, G. Starr, G. Starrett, A. Mittal, and Y. Cao, *Nano CMOS Circuit and Physical Design*, John Wiley and Sons, Inc., New York, 2004.
- C. Kothandaraman, S. K. Iyer, and S. S. Iyer, "Electrically Programmable Fuse (eFuse) Using Electromigration in Silicides," *IEEE Electron Device Lett.* **23**, 523–525 (2002).
- M. Bhushan, K. Chandrasekara, M. Ketchen, and E. Maciejewski, "Method and Apparatus for Characterizing Electronic Fuses Used to Personalize an Integrated Circuit," U.S. Patent filed (IBM Docket No. YOR920040458US1), 2005.
- S. K. H. Fung, N. Zamdmer, P. J. Oldiges, J. Sleight, A. Mocuta, M. Sherony, S.-H. Lo, R. Joshi, C. T. Chuang, I. Yang, S. Crowder, T. C. Chen, F. Assaderaghi, and G. Shahidi, "Controlling Floating-Body Effects for 0.13 μm and 0.10 μm SOI CMOS," *IEDM Tech. Digest*, pp. 231–232 (2000).
- S. Polonsky and K. A. Jenkins, "Time-Resolved Measurements of Self-Heating in SOI and Strained Silicon MOSFETS Using Photon Emission Microscopy," *IEEE Electron Device Lett.* **25**, 208–210 (2004).
- K. A. Jenkins, J. Y.-C. Sun, and J. Gautier, "Characteristics of SOI FETs Under Pulsed Conditions," *IEEE Trans. Electron Devices* **44**, 1923–1930 (1997).
- M. Ketchen, M. Bhushan, and K. A. Jenkins, "Circuit to Measure High Speed Pulse I-V Characteristics with Only DC I/Os," *Proceedings of the IEEE International SOI Conference*, 2005, pp. 77–78.
- D. J. Pearson, M. B. Ketchen, and M. Bhushan, "Technique for Rapid, In-Line Characterization of Switching History in Partially Depleted SOI Technologies," *Proceedings of the IEEE International SOI Conference*, 2004, pp. 148–150.
- M. B. Ketchen, M. Bhushan, and C. J. Anderson, "Circuit and Technique for Characterizing Switching Delay History Effects in Silicon on Insulator Logic Gates," *Rev. Sci. Instrum.* **75**, 768–771 (2004).
- M. B. Ketchen and M. Bhushan, "Anomalous History Behavior in Stacked PD SOI Gates," *Proceedings of the IEEE International SOI Conference*, 2003, pp. 168–169.
- M. B. Ketchen, M. Bhushan, and S. Bermon, "Switching Delay Variability in NMOS and PMOS PD-SOI Passgate Circuits," *Proceedings of the IEEE VLSI-TSA International Symposium on VLSI Technology*, 2005, pp. 68–69.

Received September 30, 2005; accepted for publication March 3, 2006; Internet publication June 27, 2006

Mark B. Ketchen *IBM Research Division, 2050 Rt. 52, Hopewell Junction, New York 12533 (mketchen@us.ibm.com)*. Dr. Ketchen has a B.S. degree in physics from MIT and a Ph.D. degree in physics from the University of California at Berkeley. He served for four years as an officer in the U.S. Navy, and for the last 29 years has held a variety of research and technical management positions at IBM, including serving as Director of Physical Sciences at the IBM Thomas J. Watson Research Center for several years in the 1990s. His technical expertise is in the area of microelectronic devices and measurement techniques. He currently serves as a Senior Technical Advisor to the IBM Microelectronics Division in the design, implementation, and use of advanced semiconductor test structures. Dr. Ketchen is a Fellow of the IEEE, a Fellow of the American Physical Society, a Member of the IBM Academy of Technology, and the recipient of the 1995–1996 American Institute of Physics Prize for Industrial Applications of Physics and the 1996 IEEE Morris E. Leeds Award.

Manjul Bhushan *IBM Systems and Technology Group, 2455 South Road, Poughkeepsie, New York 12601 (bhushan@us.ibm.com)*. Dr. Bhushan received a Ph.D. degree in physics from Clemson University. She has more than thirty years of experience in basic and applied research, development, and design. Prior to joining IBM in 1997, she held university, government laboratory, and industrial positions in the areas of compound semiconductor thin-film photovoltaic cells and fabrication technology for superconducting devices used in magnetic field detection and microwave applications. She was a pioneer in the development and use of chemical–mechanical polish planarization techniques for fabricating deep-submicron superconducting tunnel junctions used in high-frequency digital logic circuits and for superconducting scanning magnetometers. Dr. Bhushan now works as a Senior Technical Staff Member in the area of CMOS technology performance characterization and evaluation and the design of test structures for process monitoring and model-to-hardware correlation for the IBM Systems and Technology Group.