ON THE MODELING AND TESTING OF GATE OXIDE SHORTS IN CMOS LOGIC GATES

Hong Hao and Edward J. McCluskey
Center for Reliable Computing
Stanford University
Stanford, California 94305-4055

Abstract

The electrical and logic operation of CMOS simple logic gates in the presence of gate oxide shorts is analyzed using realistic defect models. These models reflect the resistive nature of gate oxide shorts and the difference between n- and p-channel transistors. The resistance of a short plays a central role in determining the actual circuit behavior. Faults caused by gate oxide shorts can be dependent not only on inputs to the gate containing the fault but also on other signals in the circuit, and can escape tests generated using normal TPG schemes. The stuck-at test set for a logic gate cannot guarantee to detect all transistor gate-to-source and gate-to-drain shorts in the logic gate. Gate oxide shorts in n-channel transistors affect circuit operation more severely than those in p-channel transistors do. Some limitations of present transistor-level fault modeling techniques are revealed.

Introduction

Gate oxide shorts are unintended electrical connections through the gate oxide between the gate and source, drain or channel of an MOS transistor. Gate oxide shorts have been identified as a major type of fabrication defect and, in some CMOS processes, the dominant defect [10]. They are also an important source of CMOS IC reliability problems [21]. Traditionally, gate oxide shorts have been represented by shorts between the gate and source or drain of a transistor. These shorts were first considered in [3] for testing purposes. It was found that although the effects of such shorts cannot be fully modeled at gate level, the stuck-at fault test set for a logic gate nonetheless detects all of them at the output of the logic gate containing the shorts. Transistor level test generation algorithms for these shorts were presented in [17]. Tests are generated to establish a static current path between $V_{DD}$ and $V_{DD}$ through a short, and propagate the fault effect to a primary output.

Modeling of transistor gate-to-source and gate-to-drain shorts has been considered both at gate level and at switch or transistor level. Al-Arian transformed the effects of these shorts into equivalent gate level stuck-at faults in sequential circuits [2]. Banerjee introduced a five-valued algebra to more accurately reflect the fact that intermediate voltage levels exist and the final outcome of a short could depend on the strengths of the transistors connected through the short [4]. Rajsuman found that although switch-level modeling was in agreement with circuit level analysis for NMOS circuits, the effects of many shorts in CMOS circuits could not be determined due to intermediate voltages [16]. This was also pointed out by Storey, who then proposed a transistor-level modeling method which predicts circuit logic operation by calculating the voltages at circuit nodes based on pre-calculated linear region resistances of transistors [19].

However, experimental work in recent years on the properties of CMOS gate oxide shorts, [10]-[11], [18], has shown evidence of a gap between the traditional models and what actually occurs. Specifically, a common feature in previous works is that either shorts were assumed to have no
resistance or the effects of resistance were not explicitly considered. For shorts between signal lines, which are usually metal or poly lines, Acken showed that ignoring the resistance of shorts does not affect the prediction of circuit logic operation [1]. However, for gate oxide shorts, strong evidence as discussed later indicates that their resistances are in a large range. Therefore, ignoring the resistance values of gate oxide shorts could result in misleading results. To distinguish the resistance aspect of shorts, shorts whose resistance is small enough to be ignored are referred to as hard shorts, [13], whereas shorts whose resistance cannot be ignored are referred to as resistive shorts, [4], [9]. Another assumption in previous works is that gate oxide shorts in n-channel and p-channel transistors could be modeled in the same way. However, due to special features in CMOS processes discussed later, gate oxide shorts in n-channel transistors have different properties from those in p-channel transistors and should be modeled differently, [18], [20].

The above observations imply that the traditional electrical models that represent gate oxide shorts as highly conductive gate-to-source and gate-to-drain connections in both n- and p-channel transistors need to be modified. Accordingly, previous results about circuit operation in the presence of gate oxide shorts need to be reexamined.

The impact of a short's resistance was first studied in [13] for shorts between logic nodes without feedback. It was found that, while hard shorts definitely cause incorrect logic value at one of the shorted nodes, circuit operation could be either incorrect, possibly incorrect, possibly correct or correct depending on the resistance of the short. Resistive shorts within CMOS gates were studied in [9]. Dynamic behavior of a circuit can be severely affected and drastically different for slight differences in a short's resistance. Moreover, it was found that faults caused by gate oxide shorts could depend not only on the inputs of the logic gates containing the faults but also on inputs to other gates. Tests generated using normal test generation schemes can miss targeted defects.

The purpose of this paper is to analyze the effects of gate oxide shorts in the context of logic gates by using realistic gate oxide short models which more closely reflect physical reality. It is found that the p-channel transistors can partly reduce the degradation effects due to gate oxide shorts. Pattern dependent faults caused by gate oxide shorts are studied in more detail than reported in [9]. Knowledge of pattern dependent faults due to gate oxide shorts provides us with guidelines or heuristics for test generation to avoid missing detectable faults. It is shown that the previous conclusion that the stuck-at test set for a gate could detect all gate-to-source and gate-to-drain shorts in the gate cannot be applied to real gate oxide shorts. The transistor level fault modeling technique that predicts the outcome of a short based on pre-calculated transistor resistances is shown to have some inaccuracies when applied to gate oxide shorts. Transistor level modeling also need to be modified to take into account the difference between gate oxide shorts in n- and p-transistors. Improvement in accuracy can be based on analysis of actual circuit electrical behavior.

Shorts (including gate oxide shorts) have been examined for the purpose of current or IDDQ testing, [12], [14], which has been shown to be a very effective way to detect shorts in static CMOS circuits. In this paper, we are more concerned with circuit behavior in terms of voltage levels. Since IC testing is still mostly performed by voltage sensing at circuit outputs, it is essential to understand the effects of gate oxide shorts on voltage levels in order to evaluate the effectiveness of the traditional test and fault modeling techniques for gate oxide shorts.

This paper is organized in the following way. We first discuss physical mechanisms that cause gate oxide shorts, and then present electrical models for each type of gate oxide short in CMOS circuits. Detailed analyses of gate oxide shorts are presented next, followed by summary and conclusions.
Electrical Models of Gate Oxide Shorts

A detailed account of the physical origins of gate oxide shorts is given in [18] and [22]. Basically, gate oxide shorts are caused either by defects in the gate oxide or by excessive voltage across the gate oxide. Defects in the gate oxide occur during fabrication and are random both in size and location. Breakdowns in the gate overlap regions cause shorts between the gate and source, or gate and drain of a transistor, and breakdowns in the channel region cause shorts between the gate and channel of a transistor. Large defects may cause immediate failures whereas others cause time-dependent dielectric breakdown. Transistor gate-to-source and gate-to-drain shorts can also be caused by electrostatic discharge (ESD) or electrical overstress (EOS). Possible gate oxide shorts are illustrated in Fig. 1 in a generic MOS transistor.

Figure 1: Possible gate oxide short types

Gate oxide shorts can occur during or after fabrication. Therefore, in addition to being a major yield detractor, gate oxide shorts are also an important reliability concern. Hawkins confirmed that many gate oxide shorts that do not initially cause failures could fail later [11]. Although stress screening under high operating voltage can effectively weed out many early life oxide failures, not all transistors can be adequately stressed and some defective oxides will fail during system operation [18]. Even for gate oxide shorts that do cause observable errors at circuit outputs, they can escape testing due to their peculiar properties and the inadequacy of present testing techniques, as discussed later in the paper.

Gate oxide shorts are typically formed due to excessive electric field strength at the defect site that causes Si and SiO₂ to melt, producing Si filaments that short the polysilicon gate to the Si surface. Because of the small sizes of the oxide defects, these filaments have resistance. However, the value of the resistance can be in a very large range and depends on many factors such as defect size, oxide thickness, and doping densities in the polysilicon gate and diffusion regions. Measured resistance of several shorts between the gate and source or drain in n-channel transistors in [18] ranged between 800 and 4k ohms. Due to the anticipated large variation, measurements for the resistance of gate oxide shorts in general are not available [5]. However, measurements of antifuse resistances in field programmable CMOS circuits can serve as a reference. Hamdy showed that the resistance of antifuses (caused by gate oxide breakdown under high voltage) in one CMOS process was clustered around 500 ohms with a range of 300 to 900 ohms [8]. Since antifuses are designed to achieve gate oxide breakdown with low resistance under uniform electrical programming environment, gate oxide shorts in normal CMOS circuit should have a much wider range of resistance. The measurements of other circuit parameters in [10] also suggest that resistance of shorts played a role because many shorts caused only performance degradation instead of errors. In fact, to match the experimental curves from [10], resistance values were also assumed in the models for transistor gate-to-channel shorts in [20]. The above evidence suggests that the resistance of gate oxide shorts is not small enough to be ignored.
In CMOS processes, the polysilicon gate is doped with either all n or all p type dopant in both n- and p-channel transistors [7]. We assume n doped polysilicon gate in this paper as this is generally the case in practice and is also consistent with the experiments in [10] and [18]. For gate-to-source and gate-to-drain shorts in n-channel transistors, resistive connections were observed in [18] and the electrical model is simply a resistor between the gate and source or drain of a transistor [20]. For the same shorts in p-channel transistors, however, the connection forms a pn junction since the transistor gate is n type. This was observed experimentally in [18]. We propose to model such shorts in p-channel transistors by a diode in series with a resistor. Only when the diode is forward biased can a current flow through the short. This model differs significantly from the previous simple resistor model which allows current in both directions. Figs. 2a and b show the models and symbols for gate-to-source and gate-to-drain shorts in n- and p-channel transistors, respectively.

![Models and symbols for gate-to-source and gate-to-drain shorts](image)

Figure 2: Models and symbols for gate-to-source and gate-to-drain shorts

Gate oxide shorts between the gate and channel of a transistor have different properties from those of gate-to-source and gate-to-drain shorts. For this type of shorts, electrical models matching experimental curves were developed in [20]. It was found that circuit behavior was affected only when the faulty transistor has dimensions W/L = 1. Since the vast majority of transistors in logic gates have W/L >> 1, we will ignore the effects of this type of shorts in this paper.

Analysis of Gate Oxide Shorts in CMOS Logic Gates

Detailed analysis of circuit operation in the presence of gate oxide shorts is presented in this section, using the models described above. It is assumed that only one short is present in a circuit at a time. We will analyze the effects of gate oxide shorts on the electrical and logic operation of fully complementary CMOS simple gates (NAND, NOR, INV). Figure 3 shows the gate oxide shorts considered in this paper in a two-input NAND gate and a two-input NOR gate. Our discussion will be concentrated on shorts in n-channel transistors, then the results are extended to shorts in p-channel transistors.

SPICE simulation is used as our main vehicle for experiments and verification. The transistor models used are the "fast p-type fast n-type" from the example 2 μm CMOS process in [7]. For ease of discussion, a logic gate containing a gate oxide short is a faulty gate. A gate that drives an input of another gate is a driving gate. A series of transistors in a gate that connect the gate output to $V_{DD}$ ($Gnd$) forms a pull-up (pull-down) path. A short is said to be provoked when a static current path between $V_{DD}$ and $Gnd$ is established through the short.
Resistive gate-to-source (GS) shorts were first studied in [9]. It was found that circuit operation in the presence of a GS short can be very sensitive to slight difference in resistance of the short. Delay faults, transition faults, stuck-at faults and stuck-open faults are possible faults caused by GS shorts. Of particular interest are the pattern dependent faults caused by a GS short. In this section, pattern dependent faults as well as their implications are examined in further detail.

We consider the GS short in a two-input NAND gate as shown in Fig. 4a. Input A of the faulty NAND gate is driven by a NAND gate with the same transistor sizes. The analysis in [9] showed that this short may cause: a Y slow-to-fall delay fault or transition fault, the faulty n-transistor stuck-open fault, and incorrect logic values at A. A normal test generator could generate a two-pattern test AB=(10, 11) to detect all these faults. But A = 1 can be obtained by IJ=00, 01 or 10, and a normal test generator could choose any of them. To generate a test for the GS short using the transistor level test generation algorithm in [17], any of IJB = 001, 011, or 101 could be picked. Figs. 4c and d show the SPICE simulation results for a 900-ohm short using the input sequence IJB=(000, 001, 010, 011), which form two 2-pattern tests, both with AB=(10, 11). It is expected that the faults should be detected by both tests, since they both provoke the short. However, the short causes an incorrect logic value at Y only in the second pattern pair but not in the first. Therefore the short will not be detected during a DC test if the first pattern pair is used, even though it is detectable if the second pattern pair is applied. Fig. 4b shows the truth table for the two 2-pattern tests where the outlined 1 indicates an incorrect logic value.

The circuit behavior described above can be explained at the transistor level. When IJ = 00, both pull-up transistors in the driving NAND gate are on, whereas when IJ = 10 or 01 only one pull-up is on. The equivalent pull-up resistance is smaller when IJ = 00 than when IJ = 10 or 01. When the resistance of the pull-up path in the driving gate is smaller, the static voltage at A, VA, is higher, resulting in a lower output voltage VY. However, the drastic difference in VY occurred as a result of the fact that the static gate-to-source voltage in T3, VAC, is close to the threshold voltage of T3 and thus any slight change in this voltage can result in large changes in the transistor’s resistance and consequently VY. A smaller resistance in the pull-up path of the driving gate also makes the resistance of the short relatively larger on the current path, relaxing the dynamic range of VAC. As a result, the falling transition at Y is faster. Therefore the dependence of the faulty behavior on the relative pull-up and pull-down strengths exists not only for static operation, but for dynamic response as well.

N-Channel Transistor Gate-to-Source Short

Figure 3: Gate oxide shorts in a two-input NAND and two-input NOR gate
In the above explanation, we looked at the circuit operation at transistor level. At gate level, this example means that whether there will be an error at \( Y \), the output of the gate containing the fault, could depend on signals that are not inputs to this gate. On the other hand, \( V_A \) is degraded when the short is provoked, and could cause an incorrect logic value at \( A \). A fault at \( A \) should be considered to be in the driving gate even though the physical defect is not, since the fault affects all the gates driven by \( A \). However, whether there will be an error at \( A \) depends not only on the inputs of the driving gate, which now contains the potential fault, but also on the value of \( B \), input to another gate, because \( A \) can be incorrect only when \( B = 1 \). This relation is illustrated in Fig. 5. Faults that exhibit such behavior are called pattern dependent faults, as defined below.

Figure 5: Fault at \( Y \) is dependent on \( IJ \), and fault at \( A \) is dependent on \( B \)

**DEFINITION:** A fault in a logic gate \( G \) is said to be pattern dependent if the existence of an error at the output of \( G \) depends not only on the inputs of \( G \) but also on inputs to other gates.

Pattern dependence caused by gate oxide shorts is a specific type of pattern dependence: it is confined to a depth of two gates, one driving the other. That is, faults in one gate could be
dependent on the inputs of the driving gate or the inputs of a driven gate. This is because the static current path through the short spans only two logic gates.

In the above analysis, the GS short is in a transistor whose source is not directly connected to Gnd. This case only occurs in NAND gates. A simpler case is when the source of the n-channel transistor containing the gate-to-source (GS) short is directly connected to Gnd. All n-transistor GS shorts in NOR gates and inverters belong to this category. Examples are Short 1 in the NAND gate and Shorts 1 and 3 in the NOR gate in Fig. 3. Such shorts have traditionally been considered to cause stuck-at-0 faults on the output of the driving gate. The effect of such a resistive short is that the voltage level for a logic 1 signal at the driving gate output will be degraded. If the resistance of the short is sufficiently low, a stuck-at-0 fault at the driving gate output is caused. However, for resistance in a certain range, whether the logic 1 at the driving gate output becomes 0 can depend on how many pull-up paths are on in the driving gate. The fewer pull-up paths are on, the weaker (more resistive) the pull-up is compared to the short, and the lower the voltage level at the driving gate output. Therefore, a pattern dependent stuck-at-0 can result. Pattern dependence in this case is mostly for DC considerations, since if the voltage across such a GS short is close to the threshold voltage of transistor to significantly affect the drive capability of the transistor, the logic value represented by this voltage at the driving gate output is already incorrect.

Many gate oxide shorts are not detectable because they do not cause observable errors. However, with pattern dependence, shorts that do cause observable errors can escape tests. Therefore, a test pattern for a GS short should provoke the short to the worst deviation from the defect-free operation. To guarantee the detection of a detectable GS short, a GS short should be tested with the least possible number of pull-up paths (the largest possible pull-up resistance) in the driving gate.

It should be noted that the pattern dependence as discussed above is observed only when the resistance of the short is considered. If the short in Fig. 4 were a hard short, the faulty n-transistor would never turn on and a transistor stuck-open fault would always exist. In n-transistors whose sources are directly connected to Gnd, hard shorts cause stuck-at-0 faults at the driving gate output. This shows the difference between resistive shorts and hard shorts, and demonstrates the significance of considering realistic resistance values of gate oxide shorts.

Another consequence of pattern dependence caused by GS shorts is that the stuck-at test set for a gate may not be adequate to detect all GS shorts within a CMOS gate, since the detection can depend on the inputs of other gates. In fact, even applying the exhaustive test patterns to the gate containing the fault may not detect a detectable GS short. This shows the limitations of the conclusion obtained for hard shorts in [3] that the stuck-at test set for a gate could detect all gate-to-source and gate-to-drain shorts.

In order to estimate the extent of pattern dependence due to a GS short, the same simulation on the circuit as in Fig. 4 was performed with the short assuming different resistance values. Figures 6a and b plot the static voltages at A and Y while Fig. 6c plots the propagation delays of the faulty NAND gate normalized against the defect-free delay. From the graphs, it is clear that the resistance of a short plays a central role in determining the circuit operation. When the resistance of the short is between about 500 ohms and 900 ohms, the static voltage at Y takes different logic values for the two patterns, indicating a pattern dependent stuck-at-1 fault at Y. For resistance smaller than 500 ohms, incorrect logic values occur at both A and Y for IIB=011, causing a pattern dependent stuck-at-0 at A. For resistance values greater than 900 ohms, the logic values at A and Y are the same for the two patterns. However, the propagation delay can be drastically different for resistance
between 900 ohms and about 1.3 k ohms. Therefore whether the short is detected in delay testing could depend on which two-pattern test is used. These graphs show that the pattern dependence observed in Fig. 4 is not a rare incidence and can exist for a range of resistance of a GS short.

The example in Fig. 4 reveals some limitations of transistor-level fault modeling that uses pre-calculated transistor resistances to predict the outcome of a GS short in a CMOS logic gate. In [19], the resistance of a transistor on a static current path is approximated by the transistor's linear region resistance. Voltage levels on the static current paths are then calculated based on these resistance values. The equivalent resistor networks for input $I_{11B} = 001$ and $I_{11B} = 011$ are shown in Fig. 7a and b, respectively. The linear region resistance values for the transistor models used are shown next to the resistors, and the calculated voltages in parentheses. Thus according to the calculation, the voltage levels at $Y$ for both input patterns would be interpreted as 1 and pattern dependence is not predicted. In fact, the calculation predicted that a more resistive pull-up path in the driving gate would cause a lower voltage at Y, opposite to what actually occurs. The reason for the discrepancy is that the resistance of a transistor is not fixed but rather can be drastically different for slight differences in its gate-to-source voltage. In practice, the applicability of this method to gate oxide shorts could be further limited, since the resistance of a short would be unknown but it is an essential element in determining the circuit operation.

To summarize, we observe that pattern dependent faults can be caused by all GS shorts. Pattern dependent faults exist for a certain range of a GS short's resistance. A GS short in a transistor whose source is not directly connected to $Gnd$ can cause very severe propagation delay degradation. The transistor-level fault modeling technique that uses pre-calculated transistor resistances has some inaccuracies when applied to gate oxide shorts.
P-Channel Transistor Gate-to-Source Short

GS shorts in p-channel transistors are shown in Fig. 3 as shorts 5 and 7 in both the NAND and NOR gates. We analyse a p-channel transistor as the dual of an n-channel transistor. The techniques used in the analysis of an n-channel transistor can be applied to the p-transistor if we exchange 1 and 0, pull-up and pull-down, NAND and NOR, respectively. The difference between a GS short in an n-transistor and in a p-transistor is that the GS short in a p-transistor has a pn junction diode in series with a resistor. The diode is forward biased when the transistor gate is driven by a logic 0 and there is a path from the transistor source to VDD. This is the dual condition for provoking a GS short in an n-channel transistor. So except for a forward-biased diode, a GS short in a p-transistor is the same as that in an n-transistor.

A forward-biased diode requires a certain voltage drop across it to conduct. For any voltage across a GS short in a p-transistor, only part of the voltage is dropped across the resistor. This is illustrated in Fig. 8 for a short without diode and a short with a diode. Suppose the voltage V is the same in both cases, then the current through the resistor in both cases is the same since the voltage distribution in the current path caused by the short is not disturbed. Since V1 is lower than V, R1 is lower than R. Therefore, for the same voltage drop across a GS short, the resistance of the short is smaller with the diode than without the diode. In other words, to cause the same voltage degradation on a static current path, the resistance of a GS short has to be smaller than if the diode did not exist. That is also to say, a short with the same resistance would cause more degradation in an n-transistor than in a p-transistor, given equivalent gate configurations and transistor sizes. In effect, the presence of the diode reduces the degradation caused by a GS short.

Figure 8: Effect of diode in a p-transistor GS short

Circuit operation in the presence of p-transistor GS shorts has all the properties observed for n-transistor GS shorts. The only difference is that a p-transistor GS short causes less degradation.
than an n-transistor GS short in the dual situation. SPICE simulations show that curves in Fig. 6 are shifted lower in resistance by 300 to 400 ohms for a p-transistor GS short when the dual of the circuit of Fig. 4a is considered.

**N-Channel Transistor Gate-to-Drain Short**

In an n-channel transistor, the faulty behavior due to a gate-to-drain (GD) short is more complex than that due to a GS short. The static current through a GS short when the short is provoked can only be in one direction. However, for a GD short, the static current can flow through the short in both directions, depending on how it is provoked.

We consider the GD short in the circuit shown in Fig. 9a. The short can be provoked by \( AC = 01,\) \( 10,\) or \( 11.\) The SPICE simulation results are shown in Fig. 9c and d, and the corresponding truth table is shown in Fig. 9b. We notice the difference between circuit operation due to a resistive GD short and a hard GD short. For \( R = 1.5k \) ohms, the only incorrect logic value caused by the short was at \( B \) when \( AC = 10.\) In the faulty NAND gate, the corresponding defect-free input pattern is \( BC = 00,\) which is not a pattern in the stuck-at test set for a NAND gate. This means that the stuck-at test set for the NAND gate cannot detect this short, even though the short does cause an observable error. However, when the short is a hard short, more errors are caused and the short is detected by the stuck-at test set for the NAND gate. Therefore, for the more realistic resistive shorts, the stuck-at test set for a gate cannot guarantee to detect all GD shorts in n-transistors, since an error that is expected to occur for a hard short may not actually occur for a resistive short.

Discrepancies between transistor-level modeling and SPICE simulations for resistive GD shorts are also found for the example in Fig. 9. For the transistor sizes in Fig. 9, the linear region transistor resistances are: \( R_{T1} = R_{T2} = 0.72k, R_{T3} = R_{T4} = 0.38k, \) and \( R_{T5} = R_{T6} = 1.3k. \) Transistor-level modeling correctly predicted the logic values at \( B \) and \( Y \) for \( R = 0.01k \) (hard short). However, for \( R = 1.5k,\) the prediction is only correct for \( AC = 10.\) For \( AC = 01 \) and \( 10,\) it predicted \( BY = 11 \) and \( 01,\) respectively, and is inconsistent with SPICE.

Given the fact that an n-transistor GD short can be provoked in several ways, and an error expected to occur for a hard short may not occur for a resistive short, it is important to know which of the input patterns that provoke a short will most likely cause an error. However, presently available transistor-level modeling techniques may not be reliable, due to the possible inaccuracies demonstrated in this paper. Therefore analysis at the electrical level is needed.

In Fig. 9, when input \( A = 1,\) the short is provoked regardless of the value of \( C,\) and there is a static current path consisting of \( T_6,\) the short, and \( T_1.\) However, it turns out that the voltage levels at \( B \) and \( Y \) depend to a large extent on the value of \( C. \) If \( AC = 10,\) the pull-down path in the faulty NAND gate is off and both pull-up transistors in the NAND gate are on. Whether there is an error, and whether the error appears at \( B \) or at \( Y,\) depend on the resistance of the short and the relative pull-up and pull-down strengths. When \( AC = 11,\) an additional pull-down path can be formed since the degraded voltage at \( B \) could turn on \( T_3.\) Simulations with various transistor sizes in both the driving gate and the faulty gate and different resistance of the short show that the voltage at \( B \) always stays low enough to be a logic 0. This can be explained by the fact that the pull-down paths are stronger than the single pull-up path. For this input pattern, the error (if any) will appear at \( Y.\) This means that the logic value at \( Y \) will follow that at \( B. \) When \( AC = 01,\) in addition to the pull-up path in the driving gate \( (T_3),\) the p-transistor driven by \( B \) in the faulty gate \( (T_1)\) can also be on due to degraded voltage at \( B. \) Simulations show that the voltage at \( B \) always stays high enough to be a
logic 1, because the pull-up paths are stronger than the single pull-down path. If the short results in an error, the error will appear at the faulty gate output and the logic value at Y follows that at B.

![Circuit Diagram](image)

Figure 9: SPICE simulation results for a GD short in an n-channel transistor

The above analysis has indicated a strong distinction between the cases when the pull-down path is on and when the pull-down path is off in the presence of the GD short. In a more general sense, when the gate input that drives the faulty transistor is sensitized to the faulty gate output, the error appears at the faulty gate output as a stuck-at-input fault. When the gate input is not sensitized to the faulty gate output, errors appear at either the driving gate output or the faulty gate output.

So far the discussion has concentrated on circuit behavior in relation to the inputs of the faulty gate. We now consider the effects of the inputs to the driving gate on circuit operation. These effects are found to depend on whether the gate input that drives the faulty transistor is sensitized to the faulty gate output. Pattern dependent faults can occur when this sensitization occurs, whereas pattern dependence is negligible in other cases.

As a typical example, we consider the same short in the two-input NAND gate in Fig. 9. The driving gate is now a three-input NOR gate, as shown in Fig. 10a, and the simulation results are shown in Figs. 10b and c. The transistor sizes in the NOR gate are 6/2 for the n-transistors and 3/2 for the p-transistors. Input A of the NAND gate is set to 0 with different number of pull-down paths in the NOR gate. For resistance of the short lower than about 300 ohms, the short can be considered a hard short since the two shorted gate outputs have the same logic values. For R between 300 ohms and 1k ohms, an error appears at A when J/K = 100 and B = 0 (Fig. 10b). This means that whether there is an error at A depends both on the inputs of the NAND gate and the
inputs of the NOR gate, causing a pattern dependent fault. On the other hand, when the A is sensitized to Y (Fig. 10c), the voltage levels at Y show some difference for the two input patterns, but the difference is not significant when \( V_y \) is near 2.5 V. Therefore, the pattern dependence can be ignored. The logic value at the output of the NOR gate, A, retains the correct value.

\[
\begin{array}{c}
\text{(a) NAND gate with the same GD short as in Fig. 9} \\
\text{(b) When } B = 0 \\
\text{(c) When } B = 1
\end{array}
\]

Figure 10: Pattern dependence caused by an n-transistor GD short

The above discussion can be easily extended to other n-transistor GD shorts. In general, when the gate input that drives the faulty transistor is sensitized to the faulty gate output, errors (if any) will appear at the faulty gate output as stuck-at-input faults, and pattern dependence is negligible. When the input driving the faulty transistor is not sensitized to the gate output, errors (if any) can appear at either the driving gate output or the faulty gate output, and pattern dependence can occur.

P-Channel Transistor Gate-to-Drain Short

In an n-transistor, the simple resistive connection between the gate and drain allows a static current in both directions through the short. In a p-transistor, however, static current can only flow in one direction through a GD short due to the presence of the pn junction diode. This simplifies the situations where faulty circuit operation can occur. Except for somewhat reduced degradation and fewer cases for a short to be provoked, all due to the presence of the diode, circuit operation has the same properties as discussed for n-transistor GD shorts.

The direct consequence of the pn junction is that a GD short in a p-transistor can only be provoked when the gate of the transistor is 0. This compares to a GD short in an n-transistor which can be provoked when the transistor gate is either 1 or 0. In a p-transistor, the GD short is provoked by the same patterns that provoke a GS short in the same transistor. As a consequence, tests that try to provoke a p-transistor GD short by placing a 1 at the transistor gate will be invalidated. If we consider the dual of the example in Fig. 10, then the situations in Fig. 10b will not occur because the short will not be provoked. The existence of the pn-junction diode also reduces the effect of a GD short in a p-transistor when the short is provoked, as for GS shorts in p-transistors.
Summary and Conclusions

We have analyzed CMOS circuit behavior in the presence of gate oxide shorts. The analysis is based on gate oxide short models that are more realistic and accurate than models used in previous publications. As a result, properties that would not have been noticed are observed.

The most noticeable property is that tests generated using normal test pattern generation schemes, either gate level or transistor level, may miss some faults because faults caused by gate oxide shorts can be pattern dependent. It is found that pattern dependent faults can occur for all gate-to-source and gate-to-drain shorts. For any particular gate oxide short that can cause pattern dependent faults, the dependence exists for a certain range of the short’s resistance. Pattern dependent faults caused by gate oxide shorts are localized to a depth of two logic gates. That is, a fault can be dependent on either the inputs of the driving gate or the inputs of a driven gate. Another property is that the degradation effects of a gate-to-source or gate-to-drain short in a p-channel transistor are partly reduced by the pn junction across the short. In a sense, this makes p-transistors more defect tolerant than n-transistors. Tests for gate-to-drain shorts generated using previous bidirectional short model can be invalidated for shorts in p-channel transistors.

The previous conclusion that the stuck-at test set for a logic gate could detect all gate-to-source and gate-to-drain shorts in the gate does not apply to gate oxide shorts. This is partly due to possible pattern dependent faults, and partly due to the fact that some shorts can only be detected by input patterns not in the stuck-at test set for the gate. The transistor level fault modeling technique that predicts the outcome of a short using pre-calculated transistor resistances is shown to have some inaccuracies when applied to gate oxide shorts. Transistor level modeling techniques also need to be modified to take into account the difference between gate oxide shorts in n- and p-transistors.

Circuit behavior in the presence of a gate oxide short is not easily predictable, since it depends on the resistance of the short as well as the logic gate structures in which shorts occur. Even detectable gate oxide shorts can escape detection due to pattern dependence. Such faults, called elusive faults, [15], may be a small fraction of all possible faults, but often require a large portion of engineering time to test and diagnose.

Due to the fact that many gate oxide shorts do not cause any observable errors, and the elusive nature of those that do cause observable errors, the traditional voltage sensing test technique is not effective in detecting gate oxide shorts. Effective test techniques should be able to monitor the quiescent power supply current, voltage levels at internal circuit leads, or propagation delay changes in a CMOS circuit. Therefore the IDDQ tests, the CrossCheck test technique [6], and improved delay test techniques are good candidates for effective detection of gate oxide shorts.

Acknowledgment

The authors would like to thank Piero Franco for his comments. They would also like to thank the Center for Reliable and High-Performance Computing at the University of Illinois at Urbana-Champaign for their hospitality during the preparation of this paper. This work was supported in part by the Innovative Science and Technology Office of the Strategic Defense Initiative Organization and administered through the Office of Naval Research under Contract No. N00014-85-K-0600, and in part by the National Science Foundation under Grant No. MIP-8709128.
References