ESD Protection for Mixed-Voltage I/O Using NMOS Transistors Stacked in a Cascode Configuration

Warren R. Anderson and David B. Krakauer*
Digital Equipment Corporation, Shrewsbury, MA 01545
Email: warren.anderson@digital.com, Phone: (508)841-2565
*Presently unaffiliated

Abstract - We demonstrate that NMOS transistors stacked in a cascode configuration provide robust ESD protection for mixed voltage I/O in both silicided and silicide-blocked technologies. Circuits for gate voltage modulation were added to ensure uniform finger triggering of the fully silicided device. Layout and circuit rules were developed to avoid parasitic breakdown paths.

I. Introduction

In order to follow constant field scaling requirements, the power supply voltage in many integrated circuit technologies has dropped to 3.3V or below. However, commodity ICs still operate at 5V and drive these 5V levels out of the chip. A 3.3V IC may need to receive the 5V signal, but does not need 5V on chip since driving 3.3V still meets the 5V signaling standards.

Without a 5V power supply on chip, however, this I/O configuration is incompatible with many standard ESD structures. Double-diode structures will clamp all incoming signals a diode drop above 3.3V, drawing excessive current when the IC receives 5V. Grounded-gate or gate-coupled NMOS devices would allow 5V to appear between the drain and gate, violating gate oxide reliability limits in most technologies. Process migration to shallow trench isolation (STI) increases the snapback trigger voltage of thick field oxide devices to levels where these devices are no longer effective. A Darlington-connected PNP stack has been shown to work effectively [1,2]. However, this circuit can trade off leakage at high operating temperatures for current-draw under ESD conditions.

II. Device Configuration and Characteristics

We have implemented a device, along with several layout and circuit guidelines, to provide ESD protection for such mixed-voltage I/O circuits. As shown in Figure 1, the device itself consists of a stack of two NMOS transistors in a cascode configuration, where the two devices are merged into the same active area. This configuration allows 5V to be dropped safely between the drain and source during normal operation while providing a parasitic lateral NPN for bipolar action during ESD.

With the top drain tied to the pad and the bottom source to V_{ss}, the NMOS stack is completely compati-

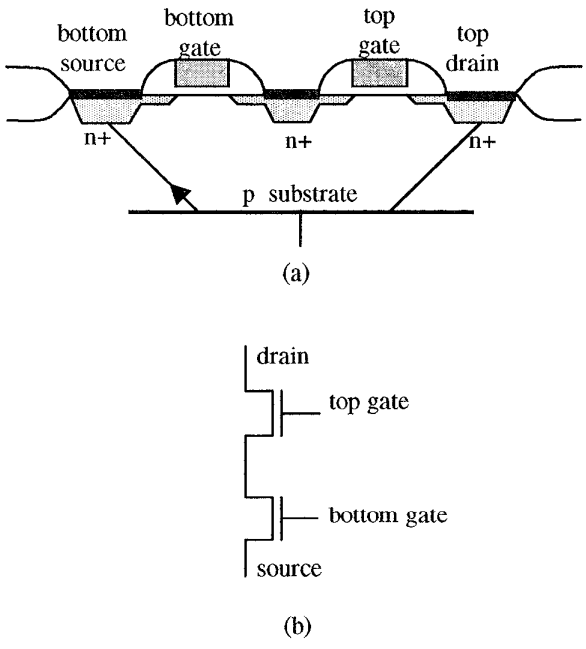


Figure 1 Stacked NMOS ESD structure in a silicided, LDD junction process: (a) cross section, (b) schematic diagram.

ble with mixed-voltage I/O. As illustrated in Figure 2, independent control of the top and bottom gates allows this device to easily meet reliability limits during normal circuit operation. The top gate is biased at or near the V_{dd} potential (3.3V), either directly or through a transistor network. With the pad at 5V, the center diffusion floats to approximately a threshold drop below V_{dd}, about 2.8V for V_{th} = 0.5V. To avoid leakage through the structure, the bottom gate must be at V_{ss}, either directly or through a transistor network. When the pad is at 5V, none of the voltages between the drain-source, gate-source, and gate-drain terminals of either transistor exceeds 3.3V. Therefore, the stack operates within dielectric and hot carrier reliability limits.

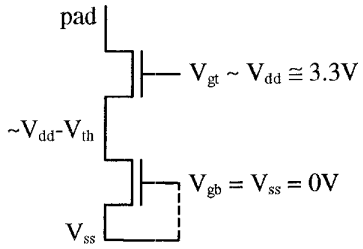


Figure 2 Biasing the stacked NMOS ESD structure under normal circuit operation conditions.

Under ESD conditions, the stacked device operates in snapback, with the bipolar effect taking place between the source of the bottom NMOS and drain of the top. These two diffusions act as bipolar emitter and collector, respectively. Their spacing determines the base width. The center diffusion floats, not participating significantly in the snapback process.

ESD clamps based on the stacked NMOS device have been used successfully in two 0.5μm technologies: one containing masked silicide block and abrupt junction steps for ESD [4], the other containing standard silicided LDD junctions with no ESD process options.

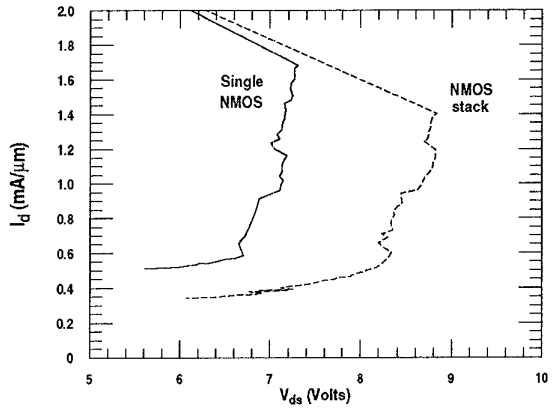


Figure 3 I-V characteristics (V_{gs}=3V) of the stacked NMOS structure and a single NMOS device in a process with fully silicided junctions.

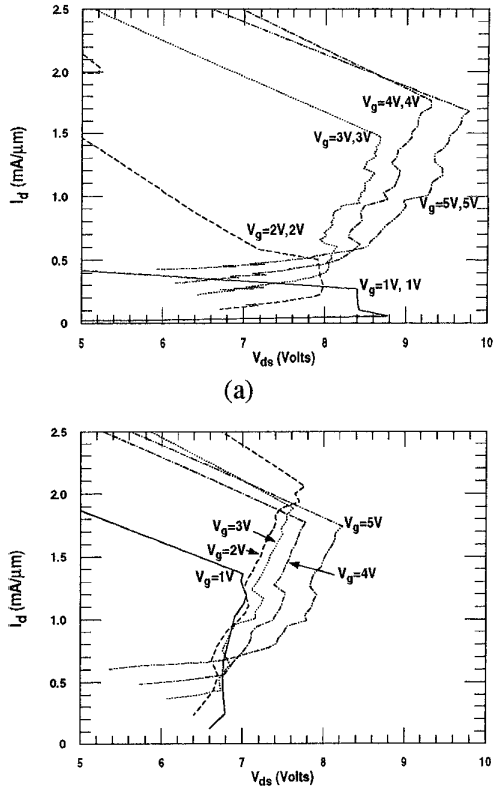


Figure 4 Gate voltage dependence of the TLP I-V in a 0.5 μm silicided process for the (a) NMOS stack and (b) single NMOS. The two V_g values in (a) signify the voltage on the top and bottom gates, respectively.

A. Silicided Technology

Figure 3 compares the snapback I-V characteristics, measured with transmission line pulsing (TLP), for a standard NMOS and the stacked NMOS device in a 0.5μm silicided technology with LDD junctions. Both structures contain only minimum length gates, with a minimum poly-poly space of 0.625μm in the stacked device. Compared to the single NMOS, the holding voltage for the stacked NMOS is 1.5V higher, a result of its roughly 1μm wider base region. The higher snapback holding voltage of the stacked NMOS also degrades its I_{t2} since the power dissipation is greater. The I_{t2} degradation is consistent with a similar power to failure for both the single and stacked NMOS.

Figure 4 compares the gate voltage dependence of the snapback I-V for a single NMOS versus a stacked

NMOS. For both devices, the gate voltage has a significant effect on the I-V characteristics. The failure current follows the same trend as the gate voltage dependence of the substrate current [3], with the maximum at approximately $V_{gs} = V_{ds}/2$. At gate voltages above this level, the field across the drain junction drops, increasing the drain voltage necessary to maintain avalanche multiplication at the drain junction. Below this level, the substrate current to maintain snapback must come from the avalanche generation induced from bipolar current more than from MOSFET drain-source current, implying a greater field at the drain. Both of these effects increase the dissipated power and therefore decrease the failure current.

Figure 4 also illustrates that the failure current I_{t2} for the stack versus single NMOS behaves differently as a function of gate voltage. The division of the top drain to bottom source voltage across the two transistors in the stack, versus the full drain-source drop in the single NMOS, explains this effect. In the stack, the voltage between the top drain and the center junction (the top device's source) controls the electric field in the top device, and therefore the channel current induced impact ionization at the top drain junction. The center junction always rises some amount above ground when both top and bottom gates are above threshold. Therefore, compared to a single NMOS, the center junction in the stack alters the drain-source voltage of the top device. The voltages on the top and bottom gates in the stack set the voltage on the center junction, therefore causing the difference in I_{t2} versus gate voltage.

The gate voltage dependence is important in protection schemes which bias the gate above threshold to ensure uniform finger triggering. This will be discussed further in Section IV-B.

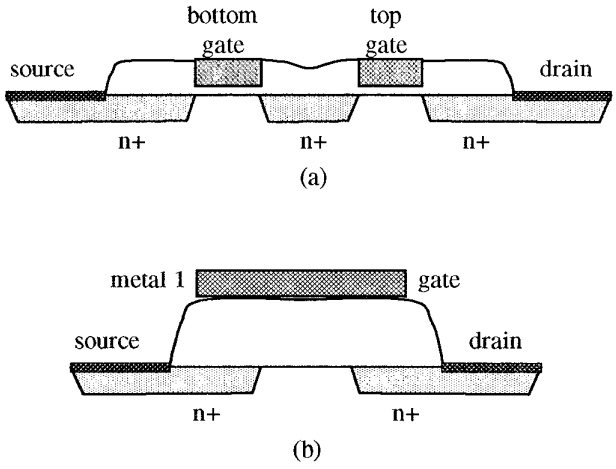


Figure 5 Device cross-sections in silicide blocked technology: (a) stacked and (b) metal 1 gated NMOS.

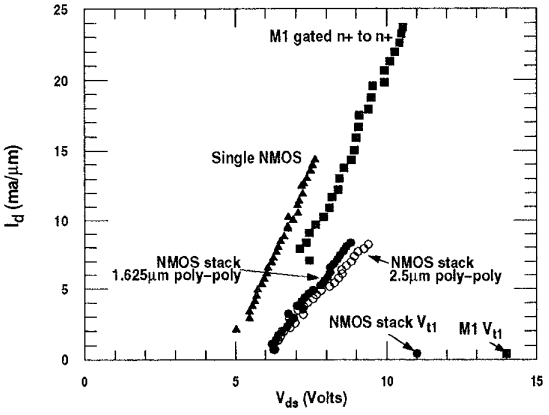


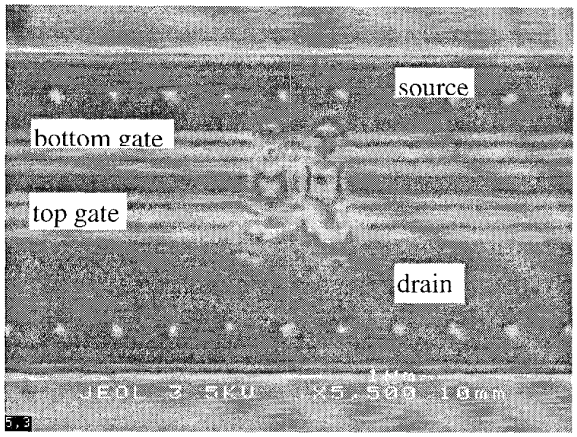
Figure 6 Grounded-gate I-V characteristics of the stacked NMOS structure, a single NMOS device, and the metal 1 gated NMOS in a process with fully silicided junctions.

B. Silicide Blocked Technology

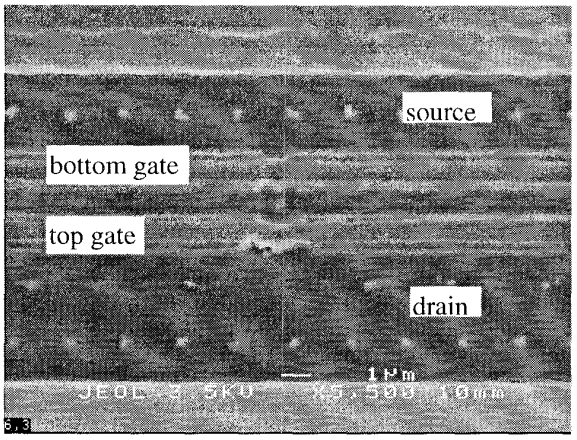
The stacked NMOS is compatible with technologies containing a silicide block process, either with, as shown in Figure 5a, or without abrupt junction implants [4]. Figure 6 shows the grounded-gate TLP I-V characteristics of the stacked NMOS device (L_{drawn} = 0.75µm for both gates) along with that of a single NMOS (L_{drawn} = 0.625µm) from a 0.5µm technology. The optional silicide-block and abrupt-junction implant was used for devices measured in Figure 6. Again, the higher snapback holding voltage and lower I_{t2} of the stacked NMOS is apparent.

Figure 6 also displays the TLP I-V characteristics of a metal-gated NMOS (L_{drawn} = 1.0µm) shown in Figure 5b. This alternative protection structure for mixed-voltage I/O in a silicide blocked technology uses masked abrupt junction implants to define the channel. It is different from thick field oxide devices since unimplanted active area, not field isolation, separates the source and drain.

The metal 1 structure of Figure 5b exhibits a 14V trigger voltage compared to 11V for the NMOS stack. Products using the metal-gated device did not consistently pass testing at 2kV HBM. The 14V needed to trigger the metal 1 NMOS could inadvertently trigger snapback in other devices tied to the pad, leading to their breakdown. Products whose metal 1 clamps were replaced with an NMOS stack of the same width consistently passed 2kV HBM, even though the I_{t2} of this device is significantly lower. This implies that a lower



(a)



(b)

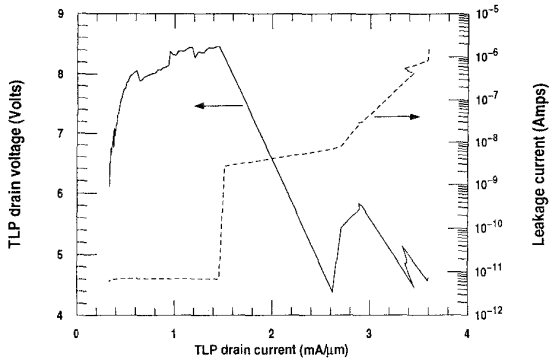
Figure 7 FA photos of failed NMOS stack devices with (a) high and (b) intermediate leakage currents.

trigger voltage is critical to providing effective ESD protection. Moreover, the mask generation complexities of the metal 1 device make it even less desirable.

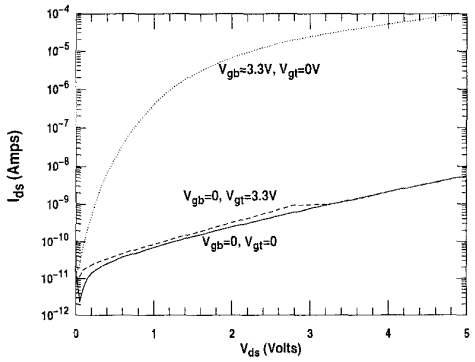
III. Clamp Device Failure Characteristics

As shown in Figure 7a, the fully-silicided NMOS stack fails by way of a short from the drain to the source through the center diffusion. At the failure point, current filamentation occurs between the source and drain, causing localized heating near the drain junction, analogous to second breakdown in a single NMOS.

Interestingly, the NMOS stack occasionally exhibits a failure across the drain to center diffusion with the bottom device still intact, as shown in Figure 7b. Additional pulses are required to short the bottom device. The TLP I-V characteristics shown in Figure 8a



(a)



(b)

Figure 8 Low-leakage failures: (a) TLP characteristics and (b) leakage I-V showing failed top NMOS device.

further demonstrate a soft failure mechanism in this device. The NMOS stack enters snapback, with a holding voltage of 8V. At the point of soft failure, the pulsed voltage drops abruptly to about 5V, the snapback voltage of a single NMOS. The TLP current abruptly increases, but the leakage with both gates grounded increases only slightly, into the nano-amp range. Visual inspection of a device which has only been pulsed to this intermediate level of failure, as shown in Figure 7b, indicates that the damage has not fully propagated to the source. The DC I-V in Figure 8b, with combinations of high and low voltages on the top and bottom gates, confirms the short between the drain and the center diffusion.

The degree of the intermediate failure phenomenon depends on process conditions. It was pronounced in about 75% of the lots tested, though it did not exist in the final process. It also has ramifications for circuit reliability. Since the top gate must typically be biased around 3.3V during circuit operational mode, the top device in the stack is on, acting to limit the voltage from the center diffusion to the bottom gate to reduce dielectric stress. When a short occurs across the top

device in the stack, the full pad voltage appears across the gate of the bottom device, exceeding dielectric stress limits. Post-ESD testing should ideally contain an I/O shorts test with the top gate low and the bottom gate high whenever possible. Unfortunately, the protection circuits used in both the silicide blocked and silicided technologies (see Figure 9 and Figure 10) did not allow this.

IV. Implementation as ESD Clamps

The ESD-related device sensitivities in the silicided versus silicide-blocked technologies required that the ESD clamp based on the NMOS stack be configured differently in each of these technologies. For the technology with the silicide block, the ballasting resistance of the non-silicided source and drain extensions allows the use of the stacked equivalent of the grounded-gate NMOS. For the fully silicided technology, the need to ensure uniform triggering of all fingers during an ESD event requires additional circuits to bias the gates of the stacked NMOS.

A. Silicide Blocked Technology

In the process with the masked silicide block and abrupt junction steps, the stacked NMOS was configured as a stand-alone clamp. Though a multi-fingered device was necessary here, no special techniques, other than symmetrically balanced layout, were necessary to achieve uniform triggering of all fingers.

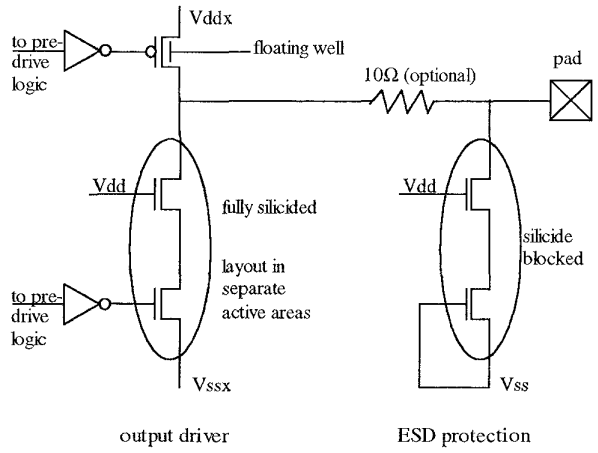


Figure 9 Schematic of stacked NMOS protection in a silicide blocked technology.

Figure 9 shows the schematic diagram of a stacked NMOS protecting an output driver in the silicide

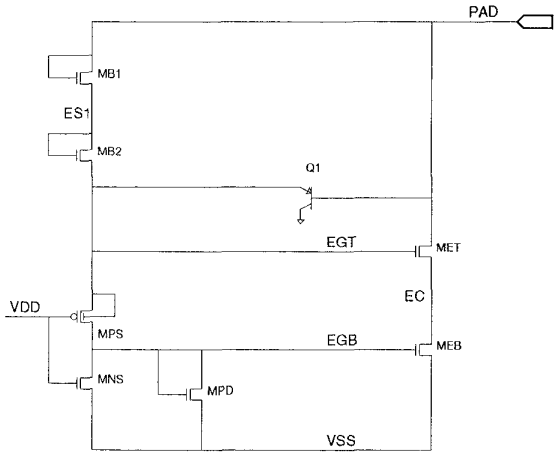


Figure 10 Circuit implementation of the stacked NMOS ESD clamp in a silicided, LDD junction process technology.

blocked technology. In order to meet gate oxide reliability requirements in normal operation, the top gate was tied to V_{dd}. The bottom gate was grounded. Under an ESD event, the protection device acts as though both gates are grounded since the de-coupling capacitance on V_{dd} causes it to float near V_{ss}.

Pins with an output stage typically contained a 10Ω series resistor, placed between the output driver and the ESD clamp. Although the resistor protects the output driver by increasing its on resistance if it conducts avalanche current, the output driver layout, as described in Section IV-C, is more critical to its ESD robustness. Likewise, the output driver layout allows it to be fully silicided.

B. Fully Silicided Technology

For the process technology where the silicide block and abrupt junction steps were not available, we also implemented the stacked NMOS structure as an isolated clamp, separate from the output drive transistors. However, a biasing network was necessary to ensure uniform triggering of all fingers [5], to bias the stack for the largest I_{t2} (see Figure 4a), and to meet voltage tolerance requirements during normal operation.

The circuit which accomplishes this is shown in Figure 10. This circuit uses the V_{dd}-referenced gate modulation concept, which distinguishes between an ESD event and a normal operation I/O transition through the voltage difference between V_{dd} and V_{ss} [6,7]. Transistors MET and MEB make up the NMOS stack device. They are each drawn 0.5μm long and 2000μm wide with a 0.625μm space between poly gates. Devices MPS and MNS act as an inverter when the pad is high,

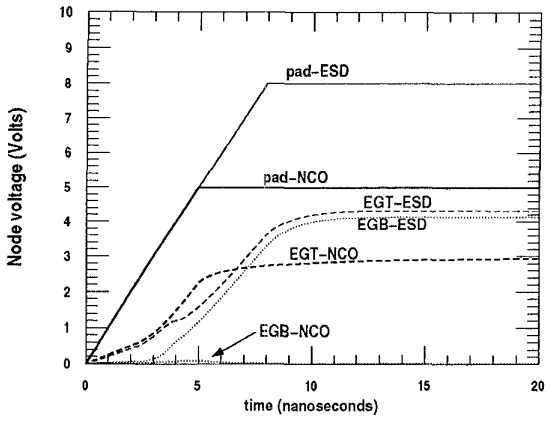


Figure 11 SPICE simulated response of the circuit of Figure 10 during ESD and I/O switching (NCO).

grounding node EGB under normal operation conditions and pulling it above V_{ss} during an ESD event. Devices MB1 and MB2 drop the voltage on the pad to bias EGT safely during normal operation. The ratio of sizes between MB1, MB2, MPS, and MPD determine the voltage on nodes EGT and EGB during ESD.

The SPICE simulated response (without NMOS snapback models) during a 1V/nsec rise on the pad to 5V in normal operation and to 8V in an ESD event is shown in Figure 11.

C. ESD Protection and Output Driver Layout Rules

The elements most sensitive to ESD damage in either configuration are the output driver and any other NMOS device with a channel-connected path between the pad and a power rail. Layout techniques applied to the ESD-sensitive devices can greatly improve their robustness. Figure 12 illustrates the layout of a multi-fingered NMOS output driver stack along with the layout of a multi-fingered ESD protection NMOS stack.

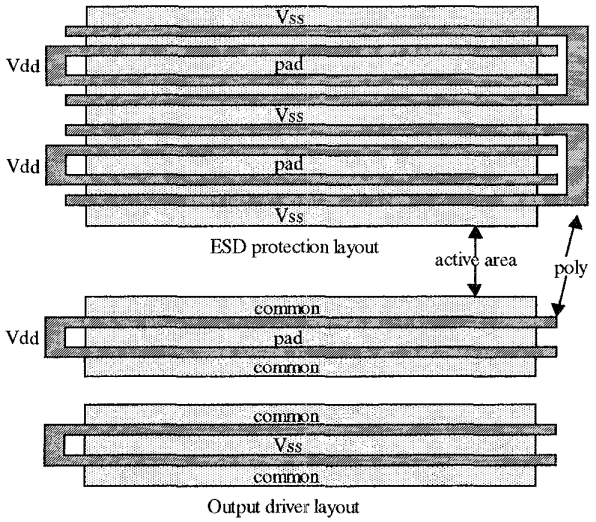


Figure 12 Stacked NMOS ESD protection (upper) layout in a single active area and stacked NMOS output driver (lower) layout in separate active areas. The metal connections between fingers are not shown.

Separating the output driver NMOS pull-down stack into different active areas significantly increases its breakdown voltage. Breakdown through series snapback of the two devices is nearly impossible. Since the common node acts as the source of the top device and the drain of the bottom, it must reach about 7V (see Figure 3) for the bottom device to fail. The resulting large source voltage on the top device forces it out of snapback (or requires an extremely large substrate current to forward bias the top source to substrate junction).

As long as both devices are of nearly equal width and have similar V_{gs} during ESD, it is also difficult to generate a snapback failure in the parasitic lateral NPN between the top drain (pad node) and the bottom source (V_{ss}). The drain junction of the top device may undergo avalanche breakdown, beginning the snapback mechanism. However, snapback across the wide base region of the lateral NPN between the pad and V_{ss} requires a large holding voltage and is therefore unlikely to occur before the ESD protection stack fails.

The robust output driver design allows the resistor in series with it (see Figure 9) to be removed if desired.

V. HBM Results

With these rules in place, the vast majority of products using either the silicide blocked or the fully silicided protection scheme met our target withstand level of 2kV HBM. While only a few products were manufactured with the silicide block option, at least 20 different product ICs using the fully silicided NMOS stack protection structures were tested. Although the NMOS stack protection eventually achieved a robust protection level on all types of I/O pins, a few of the clamp's sensitivities were only understood after some empirical lessons early in the clamp's use.

VI. Circuit Sensitivities and Failures

The higher holding voltage of the NMOS stack ESD protection requires special care in identifying and

eliminating parasitic breakdown paths. Such paths generally arise through the functional circuitry connected to the pad. A weaker path in parallel with the clamp will lower the ESD robustness of the pad.

As discussed in Sec. IV-C, it is straightforward to assess the breakdown strength of the output driver or any other similarly-configured NMOS stack. However, HBM failures in other, more obscure circuits caught us by surprise. They illustrate additional circuit and layout analysis which must be considered in using the stacked NMOS protection device.

A. Field Transistors

Field transistors where the pad node faces a power supply node across a minimum-spaced isolation region showed particular sensitivity to breakdown. This situation may arise anywhere in the pad circuit. However, output driver layout was a common cause.

The output driver layout shown in Figure 12 was robust when protected with the stacked NMOS. The alternative layout shown in Figure 13 is nearly equivalent in its function under normal circuit operation. However, the parasitic field device between the pad and V_{ss} fails under HBM test conditions.

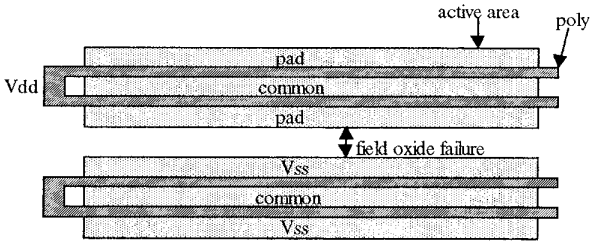


Figure 13 Stacked NMOS output driver layout sensitive to breakdown across the field region between the pad and V_{ss}.

In our process, the likelihood of failure between these two active area regions depends strongly on their spacing. At twice the minimum design rule, the field transistors stopped failing. To ensure the greatest robustness, however, the wider spacing was generally used in conjunction with the layout of Figure 12, where the common node faces itself across the field region.

B. Input Pass Gates

Some 5V-tolerant input circuits tended to fail when protected with the stacked NMOS clamp. Figure 14a shows one such circuit which failed at 500V HBM in the technology with the silicide-block option[1].

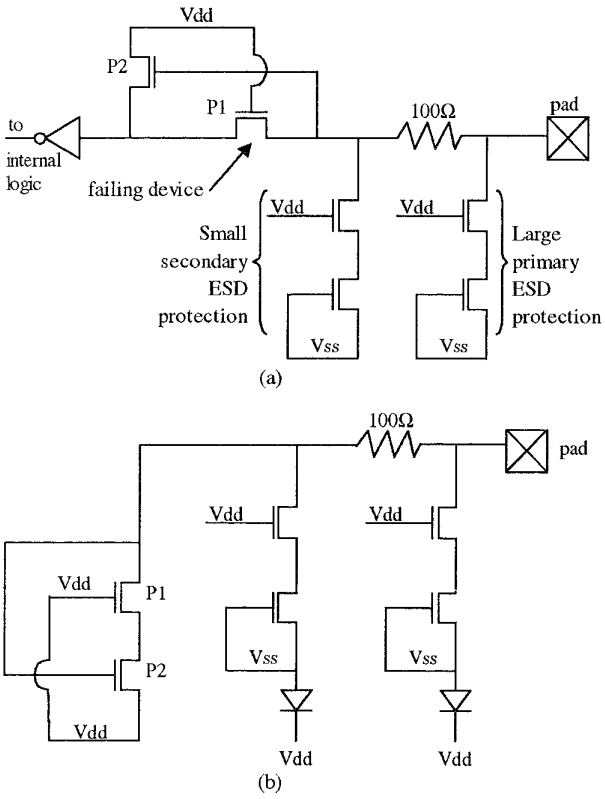


Figure 14 Input circuit which failed 500V HBM testing: (a) schematic and (b) parallel discharge paths to V_{dd}.

This circuit contains an NMOS pass gate (P1) to limit the voltage at the gate of the input inverter and a source follower (P2) to pull the input inverter's gate completely to the V_{dd} rail. The protection elements were configured in the standard pi network, with a large protection device at the pad and a small device with a series resistor for protection of the input receiver gate oxides [8,9]. The pass gate P1 fails under HBM testing.

Examining the electrical connections for a discharge from the pad to V_{dd} reveals the reason for the failure and its correction. Figure 14b unwinds the circuit to show the parallel paths between the pad and V_{dd}. We desire the current to flow through the right-most leg, through the ESD protection device to V_{ss} then through substrate-well diodes to V_{dd}. However, the current

[1] The circuit was corrected prior to the availability of products in the process without the silicide block option.

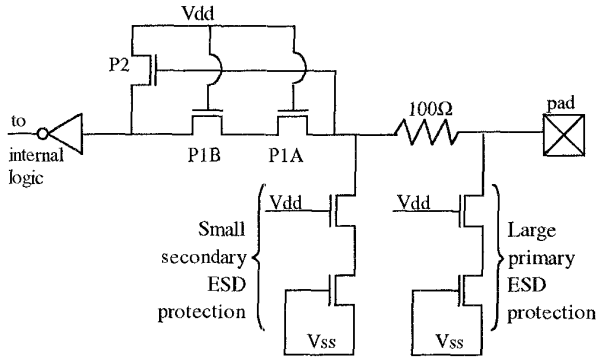


Figure 15 Correction to the input pass gate failure.

prefers to flow through the left-most leg, through the input pass gate, causing device P1 to fail.

Although both paths contain two NMOS devices in series, the difference in their gate connections accounts for the failure of the input circuit. The path through the ESD protection leg contains two NMOS devices with their gates effectively grounded (for a discharge on the pad with V_{dd} at ground). The input receiver leg contains one device with its gate effectively grounded (P1) and one device with the gate tied to the pad (P2). Device P2 will be strongly on for any reasonable voltage on the pad. The common node connecting P1 and P2 is pulled very close to ground (V_{dd}) during the discharge. The end result is an NMOS stack device protecting a single NMOS, which is clearly dangerous.

If device P1 snaps back, it sinks all of the pad current, holding the pad voltage too low to trigger the ESD protection device. According to Figure 6, the drain junction of the protection device should avalanche when the pad to substrate voltage reaches 11V. However, the drain of P1 will avalanche at this voltage as well. Similar to the case of triggering a grounded-gate, multi-fingered NMOS [5], small variations in the trigger voltage between the two devices cause one leg over the other to trigger into snapback at random.

Once device P1 snaps back, it will fail. Since device P2 is strongly on, it offers little series resistance to the snapback current in device P1. At the P1 failure current of roughly 25mA, the 100Ω resistor drops 2.5V. However, this is not enough to raise the pad voltage to the 11V needed to trigger the ESD protection device.

The general solution to this problem is to ensure that all paths from the pad to power supply rails contain a stack of NMOS transistors which remain off under ESD conditions. For the particular problem of pass gates in the input receiver, the correction, shown in Figure 15, requires two pass gates, P1A and P1B, in series. As in the case of the output driver, the pass gates must be placed in separate active areas, following the layout rules discussed in Section VI-A. Now, the discharge path from the pad to V_{dd} through the input receiver contains two NMOS devices in series, with their gates effectively grounded. We have an NMOS stack protecting an NMOS stack. This configuration was robust to over 2kV HBM.

VII. Device Scaling to a 0.25μm Technology

Figure 16 displays the TLP I-V characteristics for an NMOS stack device in a 0.25μm silicided technology. Poly gate lengths were drawn at 0.25μm with a 0.31μm space between the gates. The I_{t2} maintains its value of at least 2mA/μm, similar to the 0.5μm technology of Figure 4. I_{t2} even increases to nearly 3mA/μm at the optimal voltage of 2V on both gates, showing that the device scales well to a more advanced process technology.

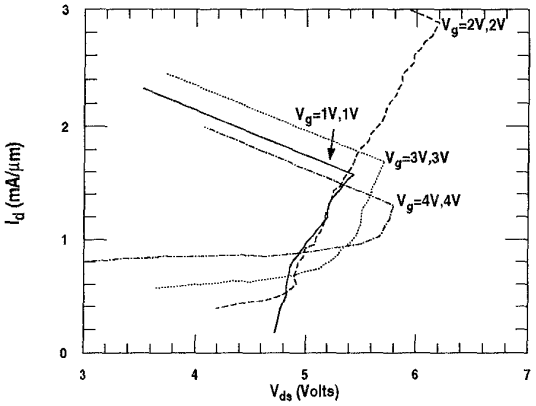


Figure 16 Gate voltage dependence of the TLP I-V in a 0.25μm silicided process

VIII. Conclusions

A stack of NMOS transistors in a cascode configuration provides robust ESD protection for ICs receiving 5V signals in a technology where 5V across the gate oxide is not allowed. Biasing the top gate of the stack allows the stack to see 5V on the pad while meeting gate oxide and hot carrier reliability requirements.

We have successfully used the NMOS stack in two silicided, abrupt junction processes, one with an optional silicide block and one with no special ESD process steps. The silicide block on the NMOS stack pro-

ess steps. The silicide block on the NMOS stack provides ballast resistance in the drain, allowing a simple biasing scheme on the gates of the stack. In the fully silicided process, a more complex biasing network brings the gates of both devices above threshold during ESD to ensure uniform snapback triggering in all fingers. We applied special layout considerations to the output driver to increase its robustness.

To avoid breakdown in other pad circuits, all paths from the pad to a power supply rail must contain a stack of NMOS transistors which will remain off during an ESD event to that rail.

Acknowledgments

We gratefully acknowledge the assistance of Kaizad Mistry in circuit design, Martin Whitcomb and Norman Leland for electrical measurements, Michael Foley and Allen LaForte for failure analysis, and Robert Woodside and Stephen Thierauf for many I/O discussions.

1. Steven H. Voldman, "ESD Protection in a Mixed Voltage Interface and Multi-Rail Disconnected Power Grid Environment in 0.50- and 0.25-µm Channel Length CMOS Technologies," *EOS/ESD Symposium Proceedings,* 1994, pp. 125-134.

2. Steven H. Voldman and Gianfranco Gerosa, "Mixed-Voltage Interface ESD Protection Circuits for Advanced Microprocessors in Shallow Trench and LOCOS Isolation CMOS Technologies," *IEDM Tech. Dig.,* 1994, pp. 277-280.

3. S. Tam et al., "Correlation Between Substrate and Gate Currents in MOSFETs," *IEEE Transactions on Electron Devices,* **ED-29,** 1982.

4. D. Krakauer and K. Mistry, "ESD Protection in a 3.3V Sub-Micron Silicided CMOS Technology," *EOS/ESD Symposium Proceedings,* 1992, pp. 250-257.

5. T.L. Polgreen and A. Chatterjee, "Improving the ESD Failure Threshold of Silicided nMOS Output Transistors by Ensuring Uniform Current Flow," *EOS/ESD Symposium Proceedings,* 1989, pp. 167-174.

6. D. Krakauer, K. Mistry, and H. Partovi, "Circuit Interactions During Electrostatic Discharge," *EOS/ESD Symposium Proceedings,* 1994, pp. 113-119.

7. Warren R. Anderson, "Circuit and Process Design Considerations for ESD Protection in Advanced CMOS Processes," Microelectron. Reliab., **37,** 1997, pp. 1087-1103.

8. R.B. Wilcox and R.E. Doucette, "The Elimination of Electrostatic Discharge Failures from Silicon Gate Logic Technologies," *EOS/ESD Symposium Proceedings,* 1985, pp. 55-58.

9. Timothy J. Maloney, "Designing MOS Inputs and Outputs To Avoid Oxide Failure in the Charged Device Model," *EOS/ESD Symposium Proceedings,* 1988, pp. 220-227.