DESIGN AND FABRICATION OF A VERTICAL POWER MOSFET WITH AN INTEGRAL TURN-OFF DRIVER

by

JOSEPH BARRY BERNSTEIN

B.S. IN ELECTRICAL ENGINEERING, UNION COLLEGE (1984)

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Sept 1986

Signature of Author ____________________________________
Department of Electrical Engineering and Computer Science
August 8, 1986

Certified By____________________________________
Martin F. Schlecht
Thesis Supervisor

Accepted By____________________________________
Arthur C. Smith
Chairman, Departmental Committee on Graduate Studies

Design and Fabrication of a Vertical Power MOSFET with an Integral Turn-OFF Driver.

by

Joseph Barry Bernstein

Submitted on August 8, 1986,87 in partial fulfillment of the requirements for the degree of Masters of Science in Electrical Engineering at the Massachusetts Institute of Technology.

ABSTRACT

The circuitry which delivers power to a computer system has not experienced the extensive miniaturization of the rest of the computer over the past decade. To reduce the size and cost of power delivery systems, the technology is being developed at MIT for increasing the operating frequency of switching converters. However, in order to achieve a significant increase in the switching frequency, the per-cycle energy dissipation in the power MOSFET must be diminished.

Losses in a power MOSFET generally result from the inability of the gate drive to adequately control the gate voltage of the switch. Parasitic capacitances interact with the resistance and inductance of a physically separate driver, resulting in a miller effect which tends to delay the turn-off and cause dissipation in the device. This effect can be minimized by reducing the capacitance associated with a vertical DMOS and eliminating the inductance associated with the gate drive.

A process has been developed, along with the necessary masks, for fabricating a vertical power DMOS with reduced capacitance for a given on-state resistance. It also contains a low voltage planar MOSFET integrated on the structure to serve as the gate driver during turn-off. The combined structure nearly eliminates the power dissipation due to the miller effect by providing a low resistance path from the gate to the source, increasing the speed of the turn-off, and reducing the power lost by nearly a factor of 5.

Thesis Supervisor	:	Martin F. Schlecht
Title	:	Asst. Prof. of Elec. Eng.

ACKNOWLEDGMENTS

This project could not have been completed without the constant guidance and assistance of my advisor, Martin Schlecht, who coached me along when I needed it, but gave me ample freedom to learn and create on my own. As for the fabrication of the devices, I am grateful to all the students and staff in the Microelectronics Laboratory at MIT who helped me master the skills required to design, lay out, and produce the devices. Particularly, I wish to thank Gary Ditmer and Lawrence Bair for having such patience with me. Furthermore, my fondest wishes go to my office mate, Khalid Rhamat, who was always there to offer a sympathetic ear. I especially wish to express my appreciation to my family and friends for their moral support and companionship whenever I needed it, and to the MIT Folk Dance club for providing a consistent source of socialization and recreation.

CONTENTS

LIST OF FIGURES

CHAPTER I.

INTRODUCTION

Over the last 10 years, the electronic circuits used in computers have undergone extensive miniaturization, but the power circuits which interface these computers to the utility have not. The power supply consequently occupies a large proportion of a typical computer's volume and it has become a major impediment to overall system size reduction.

In addition to the power circuits, large computers consuming 5-10 kW of power at 5 VDC also need a great deal of expensive and cumbersome buswork to deliver the 1-2 kA of current required. Busing the power at a higher voltage (e.g. 50 V), and therefore a lower current, would alleviate this burden. "Point-of-load" power circuits, distributed throughout the circuit boards, would then be needed to provide the final conversion to the 5 V logic level, however. A schematic of this distribution system is shown in Fig 1-1.

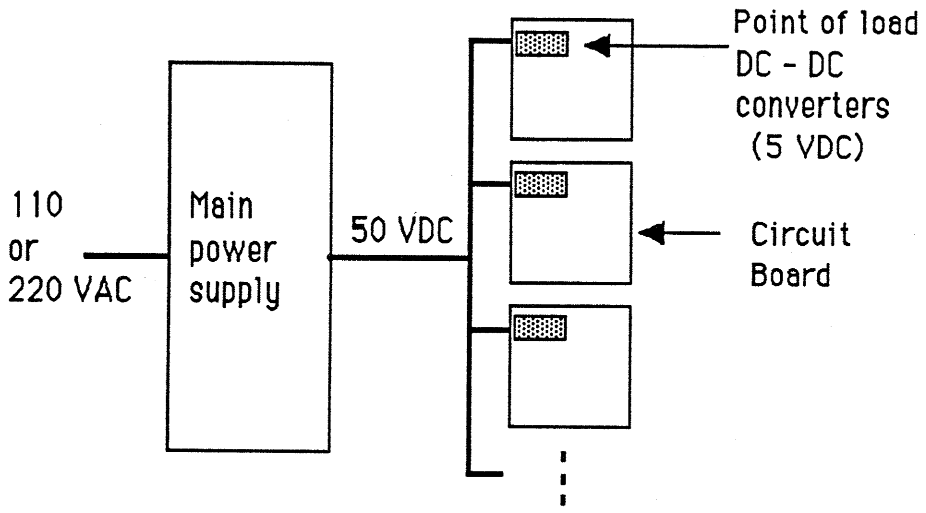


Fig. 1-1 Schematic representation of a distributed power delivery system.

In order to justify a distributed power delivery system, the point-of-load power circuits must be very small. Circuit board area is at a premium and the spacing between boards is tight, so converters need a power delivery density of 50 W/in^3 or more. Today's power electronic technology is only able to achieve densities on the order of 10 - 15 W/in^3, thus it is not practical, at present, to implement a distributed power system.

Smaller power supplies, incorporated in a distributed system, would reduce the size and the cost of a power delivery system, hence, it is important to significantly improve the power density of switching power supplies. This goal is important for all electronic systems, and it is not limited to the computer industry.

1.1. Increasing power density

Switching power supplies convert electrical power in a nearly lossless manner. Silicon power devices, switched at a high frequency, perform the basic time average conversion while inductors and capacitors remove the unwanted frequency components of the current and voltage output waveforms. Less inductance and capacitance can provide the same filter action at higher frequencies since the physical size of the energy storage elements is inversely proportional to the switching frequency. The size of the silicon devices, on the other hand, is proportional to the power requirement and not the frequency.

The silicon devices used in a typical power circuit are small compared to the inductors and capacitors. Therefore, the density of a power converter can be substantially increased by raising its switching frequency. There is historical evidence that this is true. Circuits that were operated just above the audible limit of 20 kHz with bipolar devices, for instance, delivered power at approximately 3

W/in^3. Once the power MOSFET, a majority carrier device, became available, converters could operate efficiently at around 100 KHz and deliver power at nearly 7 W/in^3.

The frequency alone does not necessarily determine the power density, however. Heat sinks and fans, which remove heat from the system, also contribute to the overall size. A point may be reached, if the frequency is raised without maintaining the efficiency, when the heat removal apparatus dominates the power circuitry. Thus, the efficiency is what limits the switching frequency for a given per-cycle dissipation. Since the switching loss increases in proportion to the frequency, the operating frequency is raised to the point where this maximum tolerable efficiency is reached. Therefore, in order to decrease the power density with increased frequency, one must first decrease the energy dissipated per switching cycle.

A common method used to reduce the switching loss takes advantage of resonant circuit elements which tailor the current and voltage waveforms for decreased dissipation. The switch can be turned ON at a time in the cycle when there are zero volts across the switch. This allows the current to increase while the voltage remains at zero, so there is no dissipation due to the turn-on transition. While the resonant approach reduces the switching loss and allows higher frequency operation, it places a larger burden on the inductors and capacitors. Consequently, larger reactive elements are needed for a resonant converter making the power density smaller for the same switching frequency.

Some small firms have shown that power converters can operate efficiently at nearly 1 MHz in a quasi-resonant circuit and achieve densities of nearly 15 W/in^3. It is believed that the density can still be improved to nearly 30 W/in^3 at this frequency if a purely switching circuit topology is employed. Such a density would be practical for the front-end converter of the computer's distributed system. To accomplish this goal, a 1 MHz power supply research program at MIT, dedicated to fabricating a high voltage front-end switching converter is underway.

In order to incorporate individual DC-DC converters on the circuit boards, however, still higher densities must be achieved. To do so requires efficient 10 MHz operation. The limitations for such high frequency operation, however, are inherent in the materials used for the inductors and capacitors, as well as in the power switches. Thus, a second project at MIT, dedicated to developing the technology for these low-end DC-DC converters with sufficient power density, is also underway. The purpose of the work presented in this document is to develop a power MOSFET with improved turn-off characteristics which will be used in both the 10 MHz resonant power supply and the 1 MHz switching power supply.

1.2. Improving the semiconductor devices

Vertical DMOS (Double-diffused MOS) transistors are used for high frequency power converters because they are very fast compared to bipolar devices. They are purely majority carrier devices and do not exhibit storage time effects. These devices can theoretically operate at switching speeds much higher than those presently used, but the gate drives are usually incapable of adequately controlling the gate voltage. These devices consequently sustain significant switching losses when operated at frequencies above 1 MHz.

One reason the gate drives are so slow is their physical separation from the power MOSFET. The parasitic inductance that exists between the gate and the driver, as well as the series resistance between the contacts and the active region of the device, limit the rate at which the gate voltage can be changed. Furthermore, energy stored in the parasitic gate capacitor is lost each time the device is switched. This energy is equal to CV^2, regardless of the resistance in series with the gate, or the speed of the switch. Hence, a larger parasitic capacitance results in more energy dissipated per cycle and a lower maximum operating frequency.

The parasitic capacitance of a vertical DMOS should be reduced to improve the switching response. The physical layout of the structure, as shown schematically in Fig. 1-2, yields depletion capacitance from the drain to source, and oxide capacitance from the gate to drain and the gate to source. The area occupied by the p-well at the source contact is not available for the on-state current flow since carriers must drift through the channel at the surface, around the p-well, and down the epilayer. Therefore, the p-well should be as narrow as possible in order to increase the conduction area and reduce the on-state resistance.

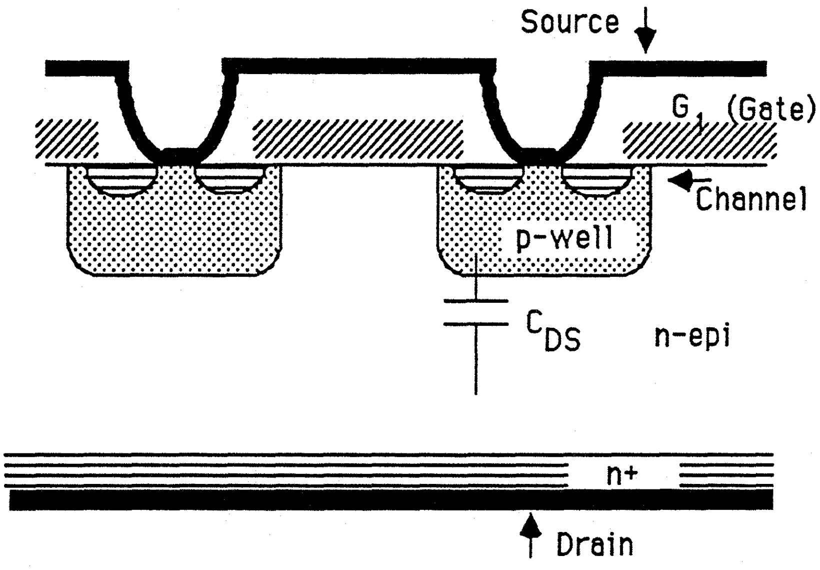


Fig. 1-2 Simplified cross section of a single DMOS cell.

The depletion capacitance of the reverse biased p-n junction is proportional to the area taken up by the p-well. This area, therefore, adds to the capacitance of the off-state device and does not add to the on-state conductance. Thus, the parasitic capacitance can be reduced for a given on-state resistance if the width of the p-well were reduced. The capacitance can be decreased in order to further reduce the dissipation in the device, to allow higher frequency operation. The device designed in this study uses high resolution photolithography in order to reduce the size of the p-well and decrease the capacitance for a given ON resistance.

The effect of using improved resolution lithography, to decrease the capacitance for a given on-state resistance, is limited. The capacitance can not actually be reduced by more than 20% to 30% for a 150 V device. The effect is even smaller for higher voltage devices, thus, simply decreasing the capacitance does not decrease the turn-off time. The turn-off response can still be improved by providing a gate driver in close enough proximity to the device itself to enhance the control imposed on the gate by the external circuitry.

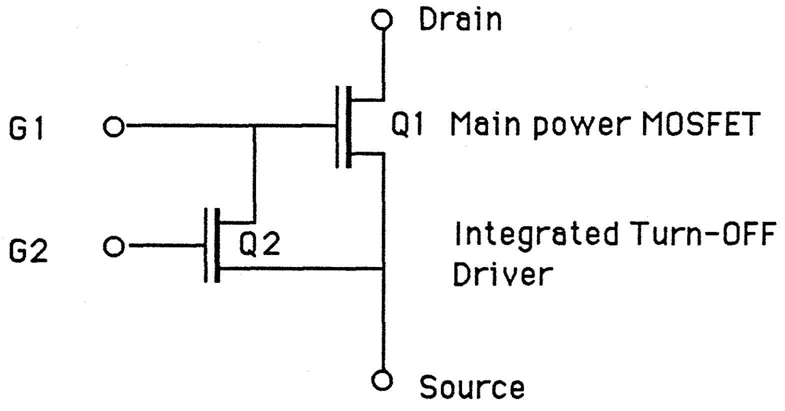


Fig. 1-3 Power MOSFET with an integrated shorting switch.

To improve the switching response, a turn-off driver will be integrated within the basic vertical power DMOS structure, that is designed for minimal capacitance given a specific on-state resistance. This structure can improve the turn-off transition time by nearly a factor of 5 over a similar device with a discrete driver. Fig. 1-3 illustrates the basic schematic of the designed device. Both gates (G1 and G2) and the source are contacted from the top of the silicon, while the drain is accessed from the n+ substrate. This device will have decreased power loss, and can be used in both the 10 MHz and the 1 MHz power supplies being developed at MIT.

The driver (Q2) physically surrounds the main device (Q1) in a way that minimizes the parasitic inductance and resistance between the two devices. The integration is accomplished in an interdigitated fashion to allow for a minimum resistance between the gate (G1) and the driver. High resolution photolithography and self-aligned diffusions also allow the designed structures to achieve a minimum possible capacitance for a required on-state resistance in order to increase the maximum operating frequency even further.

Chapter 2 introduces the theory of how an integrated driver on a DMOS with a lower capacitance can significantly decrease the switching loss for a given operating frequency. Both switching and resonant converter topologies are discussed with respect to commercially available devices and their turn-off limitations. Then the project is justified by showing how incorporating an integral turn-off driver reduces its switching loss.

The off-state capacitance of the power device depends only on its area, while the on-state resistance also depends on geometrical factors. The conduction path within the device can be optimized by using fine line lithography and an optimal layout geometry. Chapter 3 describes how this is done. The configuration of the entire structure, including the integrated shorting switch and the guard ring placement for field isolation, will also be defined.

The process sequence is then detailed in Chapter 4. The procedure to achieve the desired doping profiles, as well as specific processing details will be described.

Several difficulties arose during the process, so improvements to the fabrication process are suggested. A discussion of the masks, including specific details of the design, follows in Chapter 5, along with suggestions for how they may also be improved for future devices.

Finally, results of the fabrication and the produced devices are reported in Chapter 6. Suggestions based on the results provide a basis for continuing the investigation of increasing the maximum operating frequency of power MOSFETs.

CHAPTER II.

JUSTIFICATION

When a power MOSFET is either ON or OFF, the power it dissipates is small compared to the power being processed. During a transition between the two states, however, both a large current and a large voltage exist simultaneously. Fig. 2-1 illustrates the current and voltage waveforms of a typical power MOSFET in a square-wave switching converter. The current and voltage overlap throughout the turn-off transition (Δt_1) and the turn-on transition (Δt_2), indicating time during which the device operates in its active mode and dissipates energy.

The integral of the power dissipated in the transistor during both the turn-on and the turn-off transition is the switching energy lost per cycle. This energy, multiplied by the frequency (f), represents the switching component of the dissipated power . This component must be negligible compared to the power delivered if the converter is to be efficient. The simplest way to assure this condition is to have the total transition time ($\Delta t = \Delta t_1 + \Delta t_2$) be much less than the period ($1/f$)

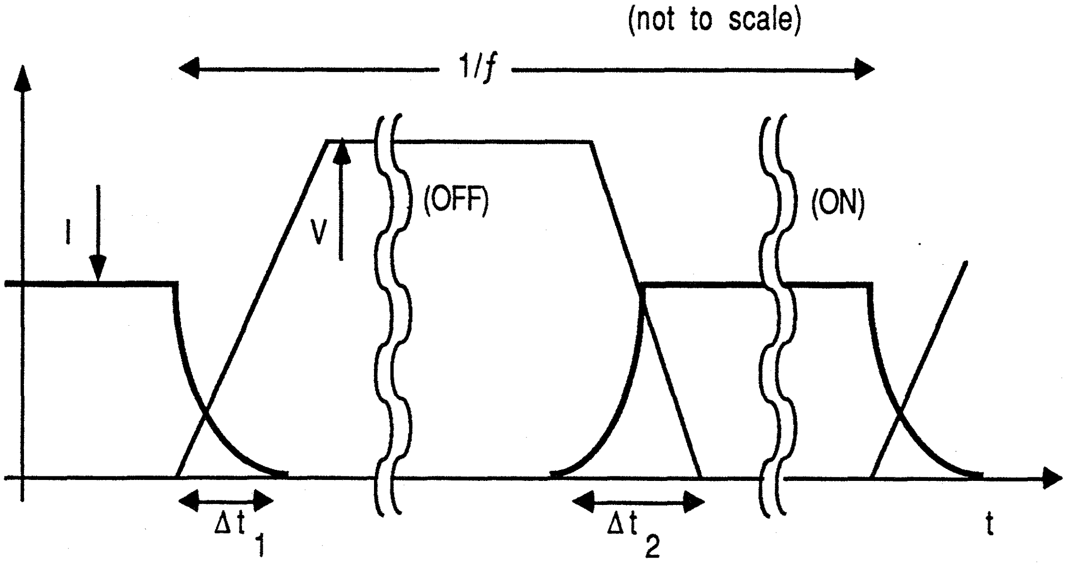


Fig. 2-1 I and V waveforms for a generalized switch in a power circuit.

2.1. Quasi-resonant approach

The switch transition dissipation can also be decreased by choosing a circuit topology that does not force the transistor's voltage and current to change simultaneously. Such a circuit is called a resonant converter because an inductor and capacitor are driven near resonance to create sinusoidal, rather than square, voltage and current waveforms. Fig. 2-2 shows a simple illustration of a resonant converter along with its associated waveforms.

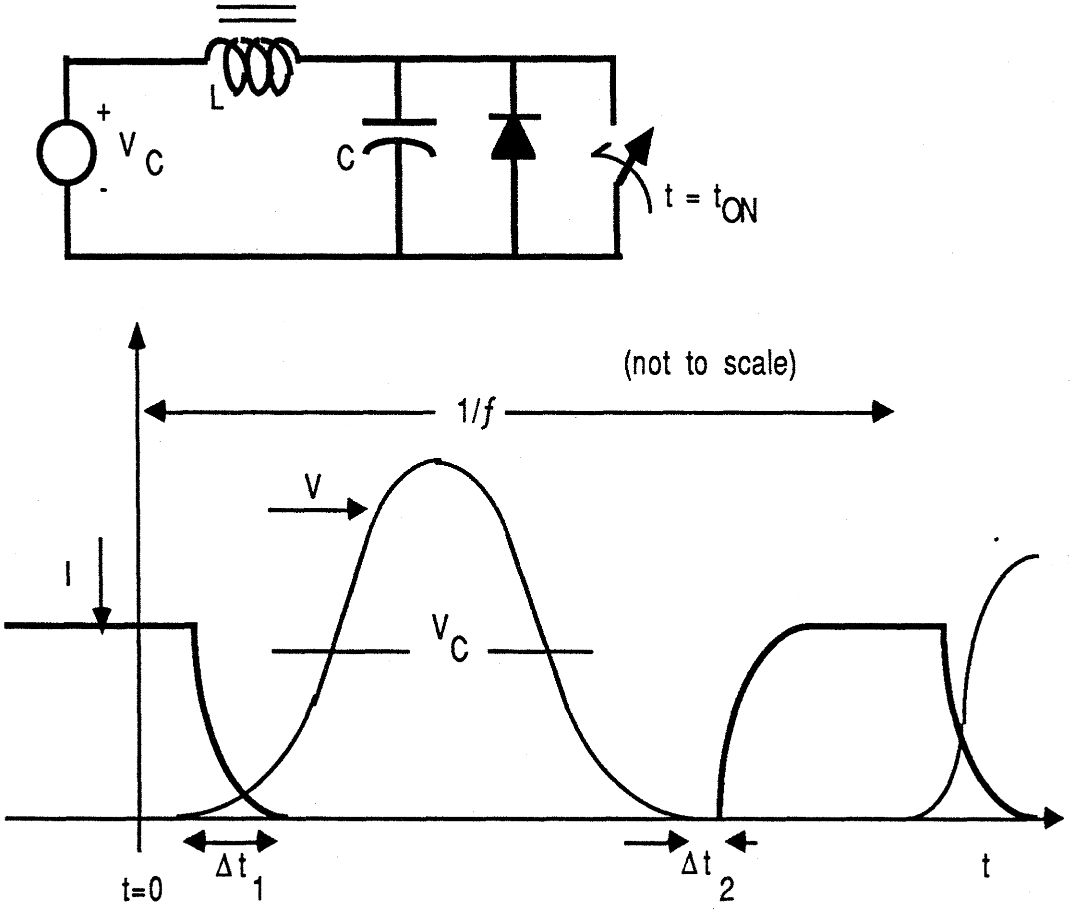


Fig. 2-2 Schematic of the basic circuit topology used for high frequency switching along with the associated I and V waveforms of the switch.

For this circuit, when the transistor is ON, it carries the full circuit current. When the device is then turned OFF at time t = 0, its current falls to zero and the inductor current instead flows into the capacitor. The voltage across the transistor rises slightly during this time because it is held down by the capacitor. After Δt_1, the switch transition is complete and all the current flows through the capacitor. Its voltage then rings in resonance with the inductor for a full cycle until it again returns to zero. The voltage never goes negative because of the diode in parallel with the switch. At this point, the switch may be turned ON with essentially zero volts across it. Once the switch is turned on, the inductor current transfers from the capacitor to the switch.

One can see from these waveforms that switching loss occurs only during the turn-off transition, when the voltage begins ramping up while the current falls off. The energy dissipated during this time (Δt_1) is less than the energy lost during the turn-off of a square wave converter (Fig. 2-1), however, because the voltage does not rise as much. This makes the switching loss of a resonant converter significantly smaller than a square wave converter, and allows operation at a much higher frequency for the same efficiency. Reducing this turn-off loss even further would allow operation at still higher frequencies.

Once the tolerable switching loss has been determined, the L and C of Fig. 2-2 are chosen to ring at the maximum frequency which yields that loss. The inductor (L) is chosen so that the ripple current is small in comparison to the circuit current. This implies that the characteristic impedance, $Z_{ckt} = \sqrt{(L/C)}$, must be much greater than π times the input impedance (V_{IN} / I_{IN}). When operating at frequencies in the 10 MHz range with an input voltage of 50 V and a current of 1 A, an inductance of 2 μH and a capacitance of 50 pf are therefore needed. This amount of capacitance is so small that it will be possible to use the transistor's junction capacitance. In fact, care must be taken to keep C this small.

For the purpose of the 10 MHz project at MIT, the following requirements must be fulfilled by the power DMOS: V_{OFF} = 150 volts, R_{ON} = 1/2 Ω. The effective capacitance of the device, which occurs as a consequence of the p-n junction from the drain to source (represented by the diode in Fig.2-2), is determined by averaging the capacitance over all values of switch voltage. For 10 MHz operation, this capacitance must be less than 50pf.

2.2. Turn-off limitations

Feedback through the arrangement of parasitic capacitors and resistors in the MOSFET tends to inhibit its turn-off transition. Fig. 2-3 is a schematic representation of the parasitic elements associated with a standard DMOS configuration that is driven by a separate driver. The current, I_0, in the external inductor of the power circuit (Fig. 2-2) flows to ground through the MOSFET. R_G represents the combined resistance of the gate material and the resistance of the gate drive circuitry.

The Miller capacitor, represented by the capacitance C_{DG}, provides feedback to the gate when the voltage across the device begins to rise. The current, I_0, has three paths through which it can flow: C_{DG}, C_{DS}, and the channel represented by the resistor R_{DS}. Current flowing through the resistor results in dissipation while current flowing through the capacitors results in stored energy that is later recovered in the second half of the L-C ringing cycle.

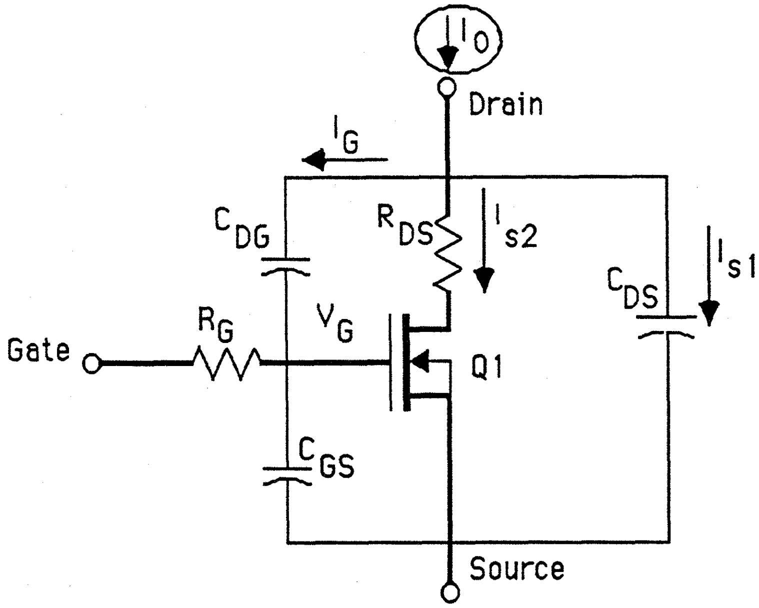


Fig. 2-3 Schematic representation of the DMOS and associated parasitic elements.

When the output at the gate driver goes to zero, the MOSFET's gate voltage decays exponentially, with the R-C time constant $R_G(C_{GS}+C_{DG})$, to the threshold voltage ($V_G = V_t$), where it remains for the miller time Δt_m. As drain current flows through the capacitors, V_{DS} rises, causing a current to flow through the gate and the series resistance (R_G). The current through the miller capacitor (C_{DG}) is limited by the gate voltage. If more current were to flow, the gate voltage would rise, more current would flow through the channel, and less current would flow through C_{DG} to the gate. Consequently, there is a feedback loop maintaining the gate at V_t while the drain voltage rises.

The drain current splits between the two capacitors, in proportion to their relative sizes ($ç = C_{DG} / C_{DS}$). Since the current in C_{DG} depends on the gate current ($I_G = V_t/R_G$), if this sum is less than the total drain current (I_0), then the remainder must flow through the channel while the drain to source voltage rises. Hence, a larger gate resistance results in a smaller gate current, forcing more current through the channel and therefore greater dissipation.

In order to avoid this effect, the gate current must be large enough to allow all the drain current to flow through the capacitors so none will be dissipated. R_G must be small enough so $V_G < V_t$ when all the drain current flows through the capacitors. The total gate resistance includes the polysilicon gate and the driver resistance, so reducing R_G involves reducing both components.

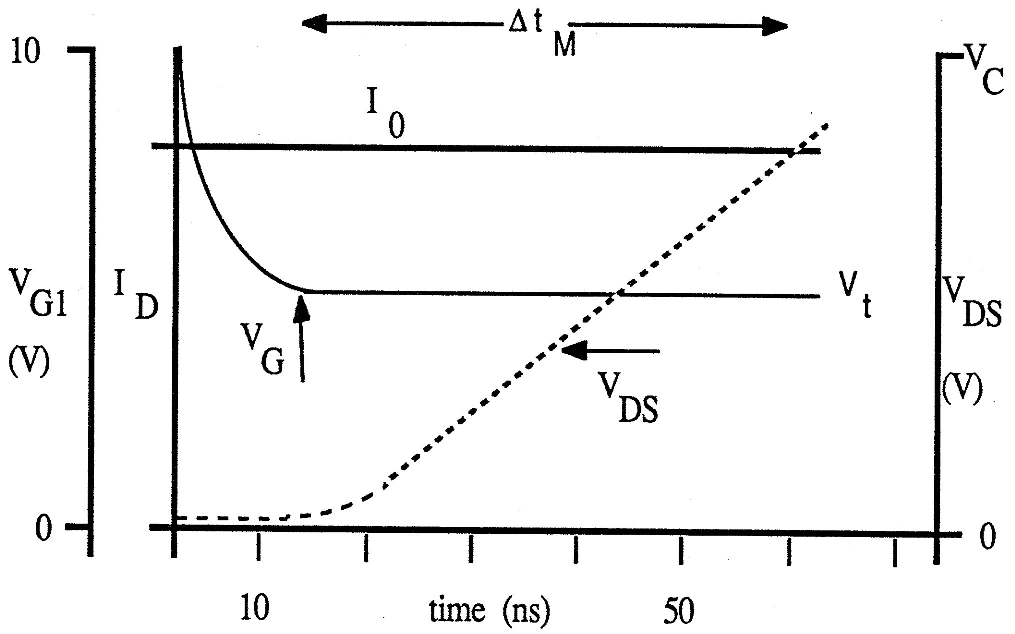


Fig. 2-4 Simulated turn-off response of the DMOS used in the case study.

Miller time is represented in Fig.2-4, by Δt_M, as the time during which the gate voltage remains constant. Voltage rises across the device with a dV/dt determined by the current through the capacitors and the combined value of C_{DG} and C_{DS}. Therefore, an average capacitance can be calculated for determining the miller time,

$$C_{avg} = (V_{OFF} - V_{ON})^{-1} \int_{V_{ON}}^{V_{OFF}} C_{DG}(V) + C_{DS}(V) \;\; dV \quad . [2\text{-}1]$$

Assuming the gate voltage remains constant at threshold (V_t), the gate current through the miller capacitor $I_G = V_t/R_G$. The ratio of I_G to I_{S1} is equal to the ratio of the capacitors, ç. The current lost to the channel resistance, therefore, is

$$I_{S2} = I_0 - I_G - I_{S1} = \underline{I_0 - (1+\text{ç})\ V_t/R_G} \qquad . \text{[2-2]}$$

Now, the miller time is calculated by assuming a linear capacitor C_{avg} with a constant current $I_0 - I_{S2} = (1+\text{ç})\ V_t/R_G$, giving a value for the miller time

$$\Delta t_M = [\, R_G C_{avg} / (1+\text{ç})V_t \,]\ [\, V_{OFF} - V_{ON} \,] \qquad . \text{[2-3]}$$

One additional characteristic of a discrete driver separated from the Gate is the appearance of a parasitic inductance. This inductance further limits the amount of cotrol one can achieve over the gate voltage. Fig. 2-5 shows the complete structure including the parasitic inductance. Changing the current through the inductor causes a parasitic impedance in series with the resistance, further enhancing the miller effect. The current can not be drawn out of the capacitor quickly, and more time is spent with current flowing through the channel during turn-off.

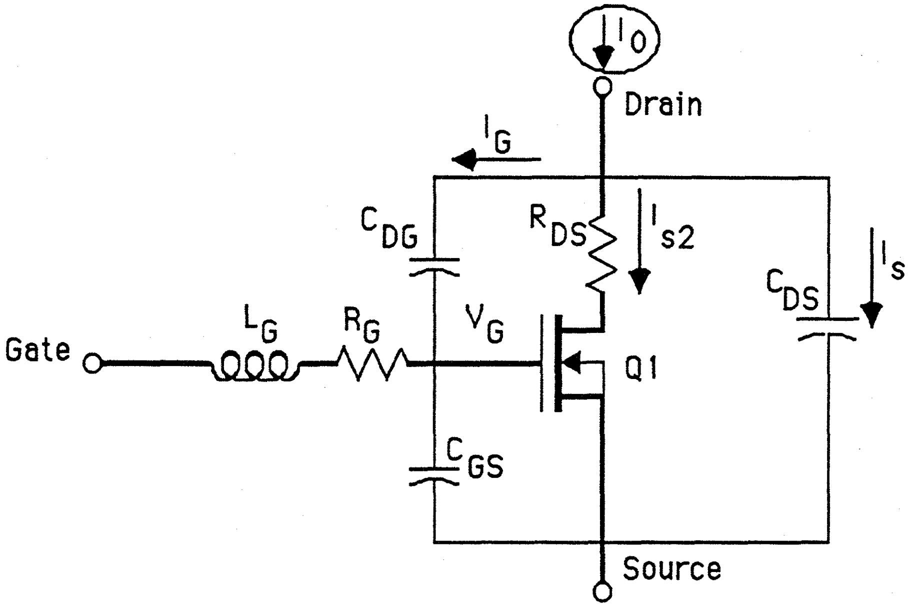


Fig. 2-5 More complete parasitic representation of the DMOS circuitry.

2.3. Integrating a turn-off driver

In an attempt to reduce the power dissipation due to the miller feedback effect of the turn-off portion of the switching cycle, we have decided to eliminate nearly all the parasitic inductance and reduce the gate resistance by integrating the turn-off gate driver on the power device. Such an integration could minimize the separation between the switch and the driver as well as reduce the gate resistance to an arbitrarily small value.

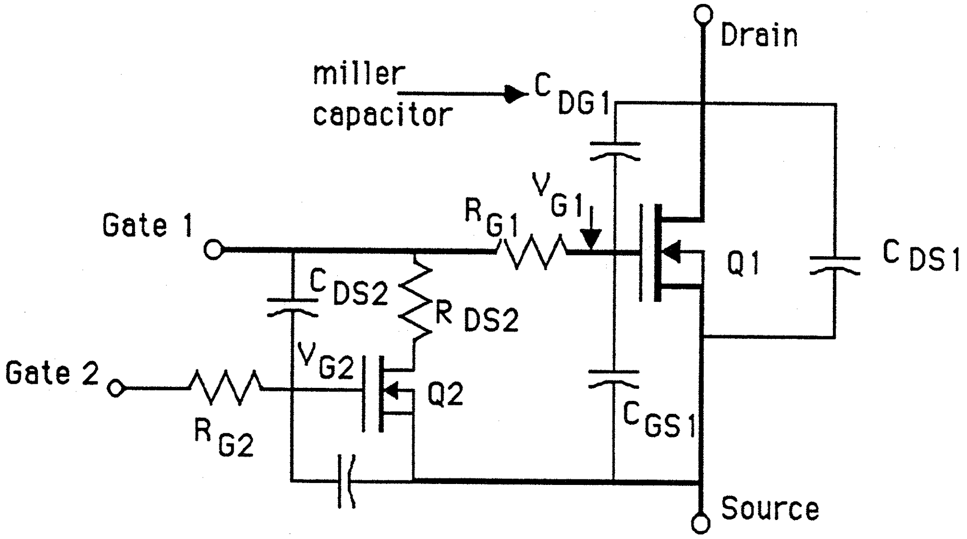


Fig. 2-6 Modified DMOS structure with an integrated turn-off driver, showing all the parasitic elements.

The gate of a typical DMOS is made of heavily doped polysilicon. The conductivity of this material is lower than 1% of aluminum and typically dominates the value of the total gate resistance (R_G). A driver integrated within the DMOS itself can be placed in a way which substantially decreases R_G. The total gate resistance can be made small enough to keep the device OFF while current flows through the miller capacitor. During the miller time, the gate voltage will be below threshold and no power will be dissipated in drain - source resistance. Fig. 2-6 shows the proposed DMOS structure with an integrated turn-off driver, including all its parasitic elements.

When the full current flows through the gate (G1), the series resistance ($R_{GT} = R_{G1} + R_{DS2}$) yields a positive voltage drop. If the total gate resistance were small enough, such that when the maximum miller current flows, the gate voltage stays below threshold ($V_G < V_t$), then the channel would remain off while the drain to source voltage rises. Consequently, the entire drain current increases the voltage across the capacitors, while no current flows through the resistive channel during the plateau time

$$\Delta t_M = (C_{avg} / I_0) (V_{OFF} - V_{ON}) \quad . \qquad [2\text{-}4]$$

No power is dissipated in the channel as long as $I_G R_{GT} < V_t$. As compared to Eq. 2–3, there is no dependence on R_G or ç for a sufficiently small series resistance. Therefore, with the value of the total series resistance

$$R_{GT} \leq (1\text{-ç}) V_t / I_0 \qquad [2\text{-}5]$$

including the driver , the gate poly, and the series impedance, the device is off, and no energy is dissipated in the channel during turn-off.

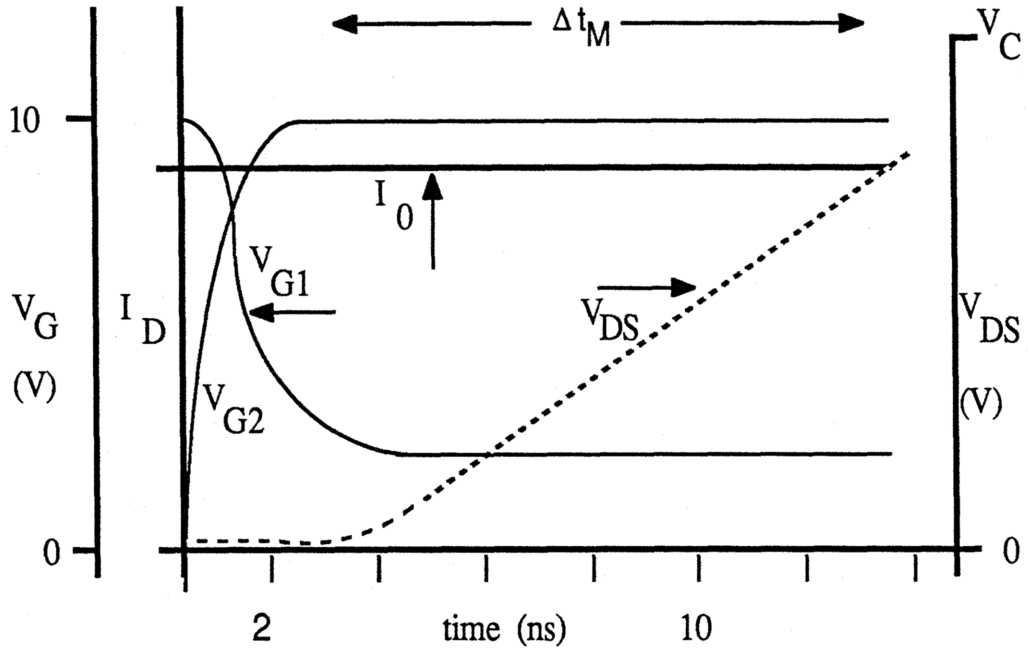


Fig. 2-7 Calculated current and voltage transient response of the modified DMOS with an integrated turn-off driver.

For the case device and circuit considered in this study, I_0 = 1A and C_{avg} = 40pf. The values for V_t = 4.5 V and ç = 0.8 as calculated from the geometry of the designed structure. The gate resistance, including the on-state resistance of the shorting switch R_{GT} = 3 Ω. Thus the criterion of eq. 2-5 is met and the power dissipation, due to the miller feedback effect, is eliminated. Fig. 2-7 shows the resulting waveform for a transition from 1 to 150 V, with the turn-off time calculated from eq. 2-4. Δt_M = 10 nS, which is an improvement by almost a factor of 5 over the same structure with a discrete driver.

Chapter III.

DESIGN DETAILS

The structure of a basic vertical n-channel power MOSFET is shown in Fig. 3-1. A lightly doped epitaxially grown n-type layer of silicon on top of a low resistivity n+ substrate forms the drain of the device, and a patterned n+ polysilicon on a 1000Å oxide forms the gate. The source and p-well diffusions are implanted through openings in the gate material, assuring their alignment to the gate edges. The channel forms in the p-type material just beneath the gate, between the n+ source and the n- epi-layer.

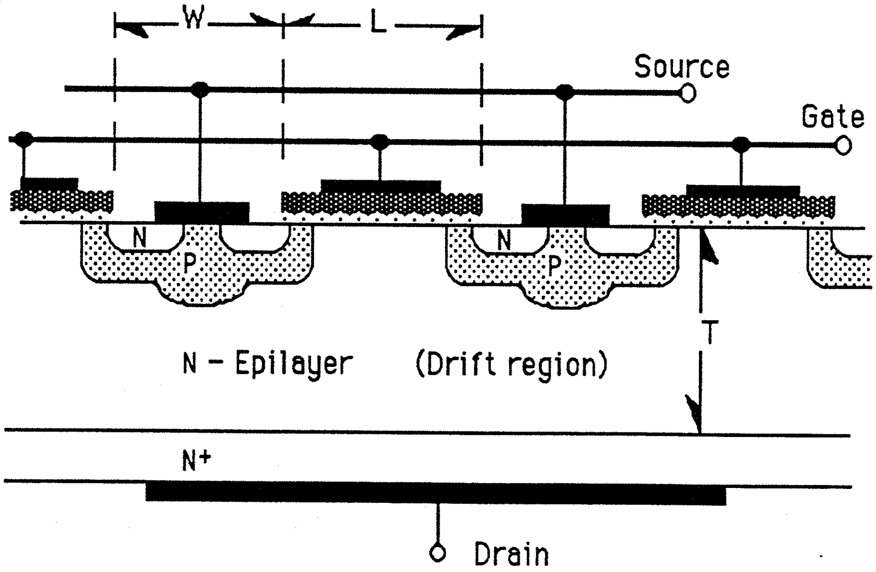
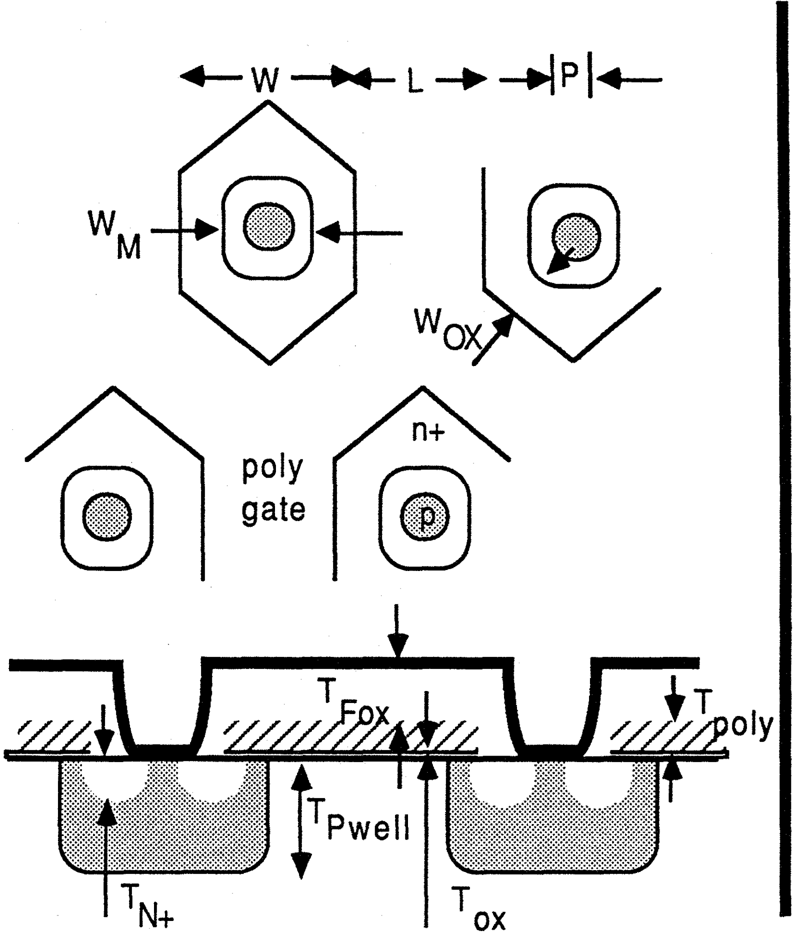


Fig. 3-1 Cross section of a typical Vertical DMOS showing the dimensions of the gate (L), the source well opening (W), and the epi thickness (T).

The p-wells, which contain the n+ sources, are arranged in a matrix as seen from the top of the device in Fig. 3-2. The individual sources are interconnected with a continuous layer of metal that is deposited over the surface of the device. This metal is electrically isolated from the gate by a 1.5 μm thick layer of oxide. The metal also shorts the source (n+) regions with the p-wells at every contact in order to avoid the detrimental effects of a floating substrate. To provide a low resistance path from the channel to the contact, the center of the p-well is heavily doped.

The channel region is self aligned to the gate by diffusing the p-well through openings in the poly. This approach minimizes the gate's overlap of the source, and therefore keeps the gate - source capacitance small. It also allows the design for the gate to be built into the same mask as the one used to incorporate the p-well, eliminating the need for alignment tolerances in the mask which defines the gate, keeping the dimensions small, and the fabrication process simple.



W =	20 μm
L =	14.5 μm
P =	9.3 μm
W_{OX} =	12.7 μm
W_M =	3.7 μm
T_{Fox} =	1.5 μm
T_{poly} =	1.5 μm
T_{Pwell} =	4 μm
T_{N+} =	1.2 μm
T_{ox} =	975 Å

Fig. 3-2 Some typical dimensions for a 200 V vertical power DMOS.

The p+ diffusion in the center of the source well is not typically self aligned in commercial power MOSFETs. This diffusion is instead introduced early in the fabrication sequence, before the gate poly deposition. The openings for the source diffusions through this poly must therefore be aligned to the deep p+ regions. As a result, the design tolerances for the process steps that follow must be large enough to accommodate the spread of the p+ region. The well width (W) should be as small as possible, however, its minimum dimension is limited by the lateral diffusion of the p+ diffusion and the subsequent alignment tolerances.

The individual source wells are typically arranged as hexagons or squares within the polysilicon gate matrx. The source forms an annulus within the p-well, leaving the center open for the metal to contact the p+ region. The arrangement of these source wells within the poly determines the device geometry. Early investigations have shown that there is no significant difference in on-state resistance between a hexagonal geometry (Fig. 3-2), and a square geometry (Fig. 3-3).[4]

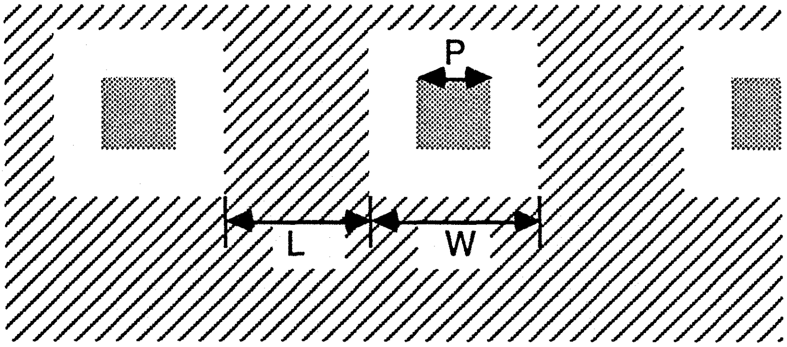


Fig. 3-3 Detail of the first mask layer defining the double diffusion within the region of side W, and the p+ contact made through the region marked P.

Hexagonal structures, on the other hand, have later been shown to give a slightly lower resistance for a given area than square ones. The effect, however, is second order (< 5%) and only becomes apparent with high current densities.[5] Also, the optimal spacing between well widths (L) is independent of the geometry (square or hexagonal), so a square geometry will suffice for investigating the optimal values of L for a given W. It is also easier to fabricate masks for the square geometry with the CAD facility at MIT.

The chip area used by the p-wells is not fully available for current to flow down the epi-layer. The electrons flow from the source, through the channel, and down to the n+ drain. The current spreads out under the p-wells at approximately 45° angles as shown in Fig 3-4. Narrower p-wells therefore increase the area available for conduction and provide a lower resistance for the same amount of area. The dimension of the source well, therefore, must be as small as possible to maximize the ON conduction for a given total device area.

The significance of this effect depends on the thickness of the epi as compared to the width of the p-well. Since the width (W) is made as narrow as lithography allows and the epi thickness depends on the off-state voltage of the device, the on-state resistance varies much more with the p-well width in lower voltage devices than in higher voltage devices. For example, the depth of the p-well is approximately 4 μm, and the width is 26 μm in a common commercial device. In a 150 volt device the epi is about 15 μm thick, so the current paths of adjacent cells do not even meet at the n+ substrate. Consequently, the on-state resistance of devices designed to withstand voltages in this range or lower depends greatly on the p-well width (W).

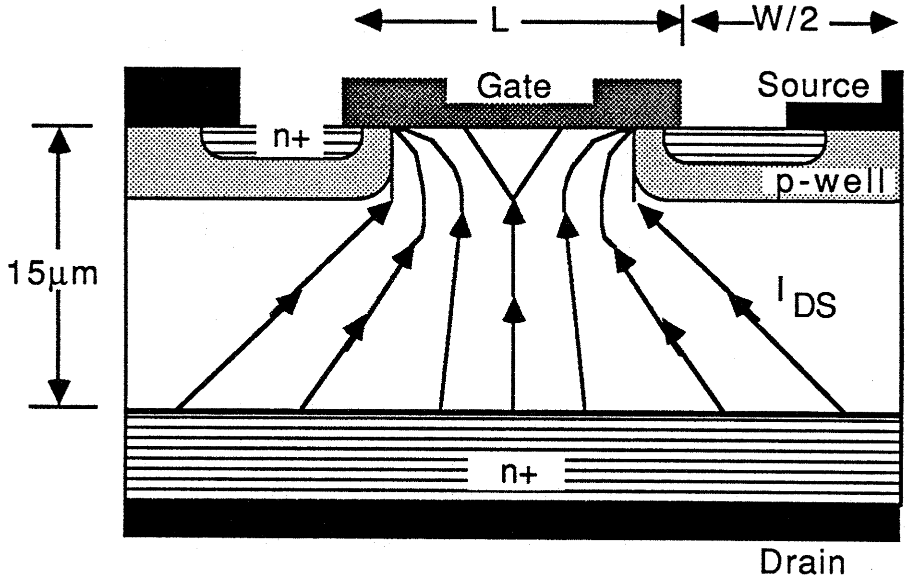


Fig. 3-4 Cross section showing the current flow path of a typical 150 V DMOS.

When a sufficiently positive gate voltage is applied with respect to the source, the p–region beneath the gate inverts and forms a channel which turns the device ON. The surface of the n–region between the p-wells also goes into accumulation with the same gate voltage. The surface of the silicon thus becomes more conductive with the accumulated charge, allowing the current to spread out between the p-wells at the surface (Fig. 3-4). However, this effect is limited. Once the lateral surface conductance becomes comparable to the vertical conductance of the bulk epi, the current no longer spreads out, but flows downward instead.

Consequently, for a given epi thickness (hence a particular breakdown voltage), there is a particular p-well separation (L) which yields a maximum use of the area between the p-wells. This value should also be independent of W since it is a measure of the effectiveness of the current spreading. One can model the spreading resistance with a two dimensional model to determine an optimal value of L for a given breakdown voltage. Hower et. al.[5] has investigated the optimal p-well spacing for both a 120 V and a 500 V device. Their results confirm the geometries used in the 200 V commercial device shown in Fig. 3-2.

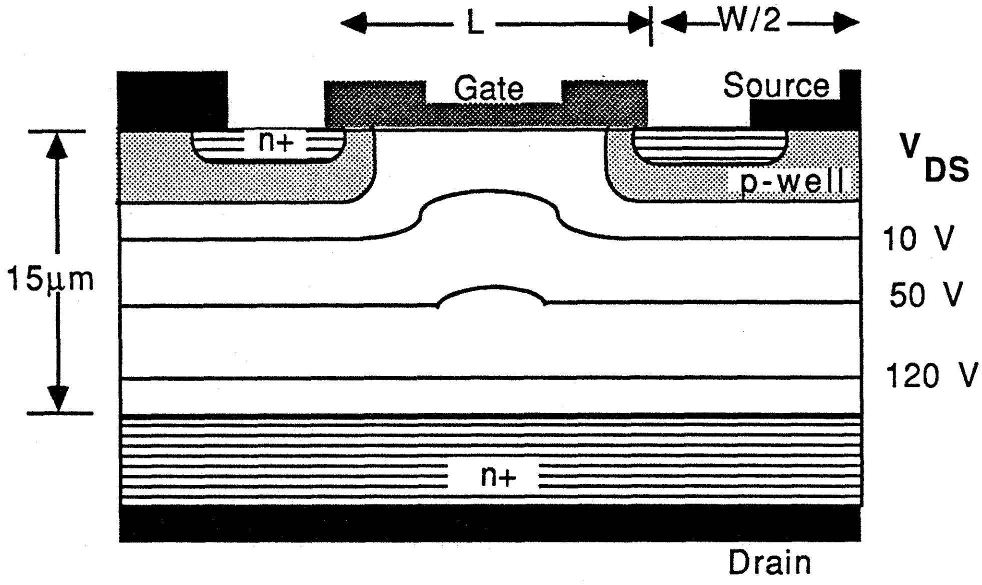


Fig. 3-5 Cross section of an OFF device showing the depletion depth with various applied drain to source voltages

When the power MOSFET is off, on the other hand, a depletion region forms within the epi-layer and spreads across the device area. For voltages as low as 10 V, this depletion layer width is nearly independent of the source well geometry, as shown in Fig. 3-5. The capacitance caused by the depletion of the OFF device is thus nearly independent of the geometry and approximately proportional to the total device area. Consequently, if a given area is chosen for a desired on-state resistance, the off-state capacitance will be nearly directly related to it, independent of the p-well geometry. If W were decreased to yield a lower resistance, less area is needed, so the off-state capacitance will be smaller for the same desired ON resistance.

Some typical dimensions for a 150 V power DMOS are shown in Fig. 3-2. These values are shown for a hexagonal geometry, but should suffice for a "first guess" with a square geometry. The epi-layer is 2 Ω-cm n-type material with a thickness of about 15 μm, which will withstand the full reverse voltage[1] as long as field guard rings provide sufficient termination at the device edges.

The field guard rings surround the active device area as shown from the top and in cross section in Fig. 3-6. These p-type rings help spread out the equipotentials near the device edges to avoid premature breakdown. The rings are usually introduced with the p+ diffusion, early in the fabrication process, before the poly is deposited. On top of the guard rings, a thick isolation oxide must be grown or deposited so metal or poly from the main device will not interfere with the field termination.

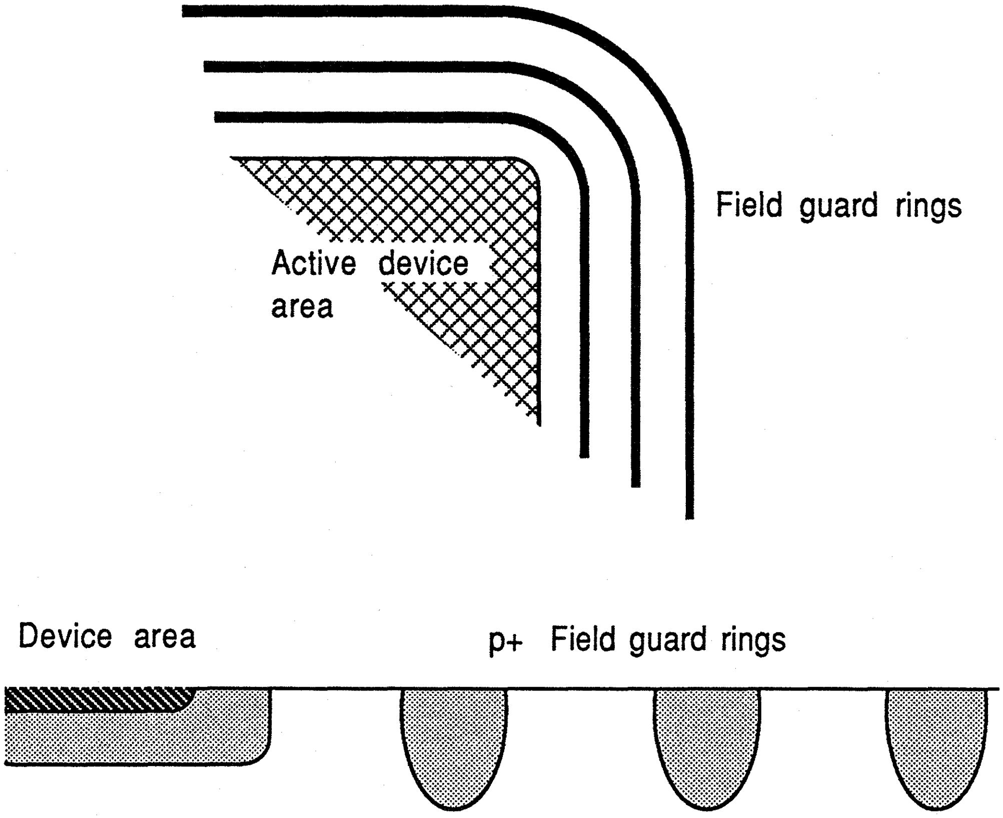


Fig. 3-6 Top view and cross section of p-type guard rings surrounding the active device area which provide field termination at the edges.

The following two sections of this chapter describe the way the process and masks are designed to reduce the capacitance for a given on-state resistance, and how a secondary lateral MOSFET integrated on the main device structure is a turn-off driver. Some of the design considerations which lead to the fabrication sequence and mask set are discussed. Also, the structure which has been fabricated at MIT is detailed. Finally, the last section provides calculated resistance and capacitance values for the completed structure.

3.1. Decreasing the on-state resistance

One can achieve a source region with the smallest possible W by having both a p-well that is self-aligned to the gate and a p+ contact region that is self-aligned to the center of the n+ annulus. The first mask, which defines the gate poly, the p-well, and the n+ source, can also define the opening in the center of the well as shown in Fig 3-7. This process takes advantage of the selectivity by which silicon can be etched with respect to oxide in a controlled plasma. The center region of width P is removed after the double diffusion of Boron (p-well dopant) and Arsenic (n+ dopant) to define the implant for the p+ contact.

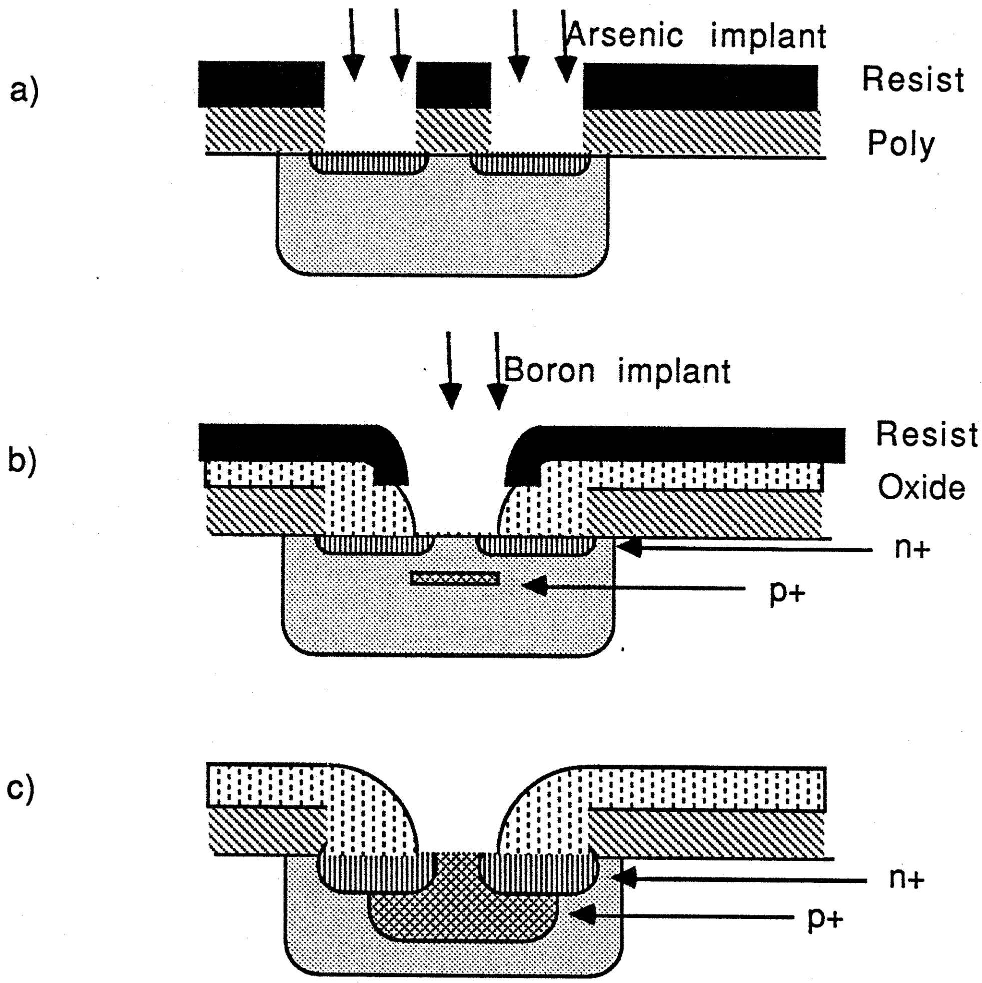


Fig. 3-7 cross sections showing double self-aligned process. a) The first mask patterns the poly for the p-well diffusion and the n+ implant. b) The poly at the center of the region is plasma etched and the heavy boron dose is implanted. c) The final diffusion step forms the junctions.

The heavy p+ Boron must diffuse beneath the n+ region, so a high temperature step follows that final implant. However, the n+ diffusion can not be allowed to diffuse more than 1 µm under the gate, otherwise the channel will be too narrow. At lower temperatures, the diffusion constant of Boron is higher than that of Arsenic[2], so the n+ region will not diffuse as far as the p+ region. Thus, a long diffusion at a lower temperature after both the n+ and p+ implants have been made will allow the Boron to diffuse farther than the Arsenic. The p+ region is more lightly doped than the n+ region which insures that the n+ region will contact the source metal. The Arsenic profile is what limits this final diffusion, thus, the final diffusion step should be adjusted for the 1µm junction between the n+ and the p-well, without concern for the p+ profile. Also, there should be no high temperature processing steps between the n+ and the p+ implants to insure a maximum diffusion of the p+ "plug" region.

Along with the implant for the source double diffusion, guard rings are defined for field termination at the edges of the devices. The p-well diffusions, used for the p-region of the "rings", consequently have n+ at the surface (Fig 3-8). This occurs because there is no masking step between the p-well diffusion and the n+ implant. Since they basically provide equipotential regions around the active device area in the OFF state, there is no trouble with isolated n+ regions at the surface. These diffusions are also aligned to the rest of the device since the first mask defines all the structures. A plasma can etch poly in a selective and anisotropic fashion giving straight sidewalls, accurately transferting the mask pattern and stopping at the underlying oxide.

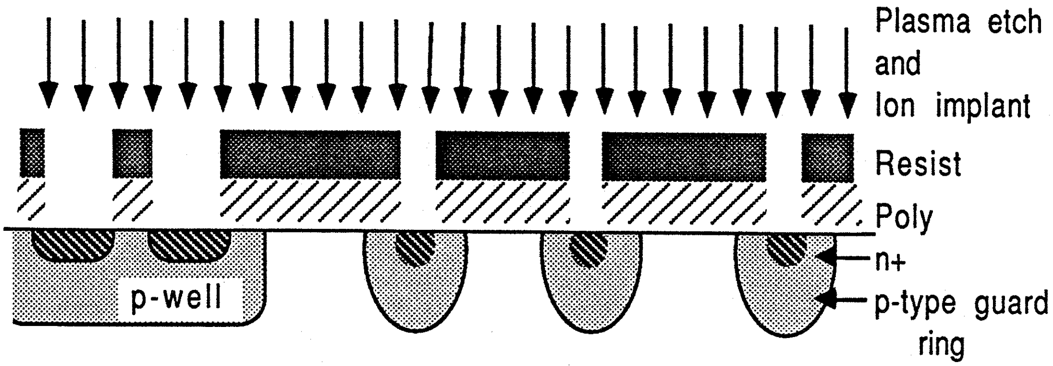


Fig. 3-8 Double diffusion and guard ring placement, defined with the first mask.

The best way to determine the optimum dimensions of the power MOSFET is to fabricate a matrix of devices on the same chip and compare their characterisics on a one-to-one basis. One of the devices with the smallest W should have the lowest on-state resistance, but it is not clear what the optimal value of L will be for that well width. Secondly, it is not clear how reliably the devices with the smallest W can be produced. Hence, in order to determine the geometry which consistantly yields the lowest ON resistance for a given area of silicon, a matrix of devices with various values of W and L have been chosen around the values which were determined as optimal from two dimensional models.[4-8] Fig. 3-9 shows the layout of the chip with appropriately labeled rows and columns.

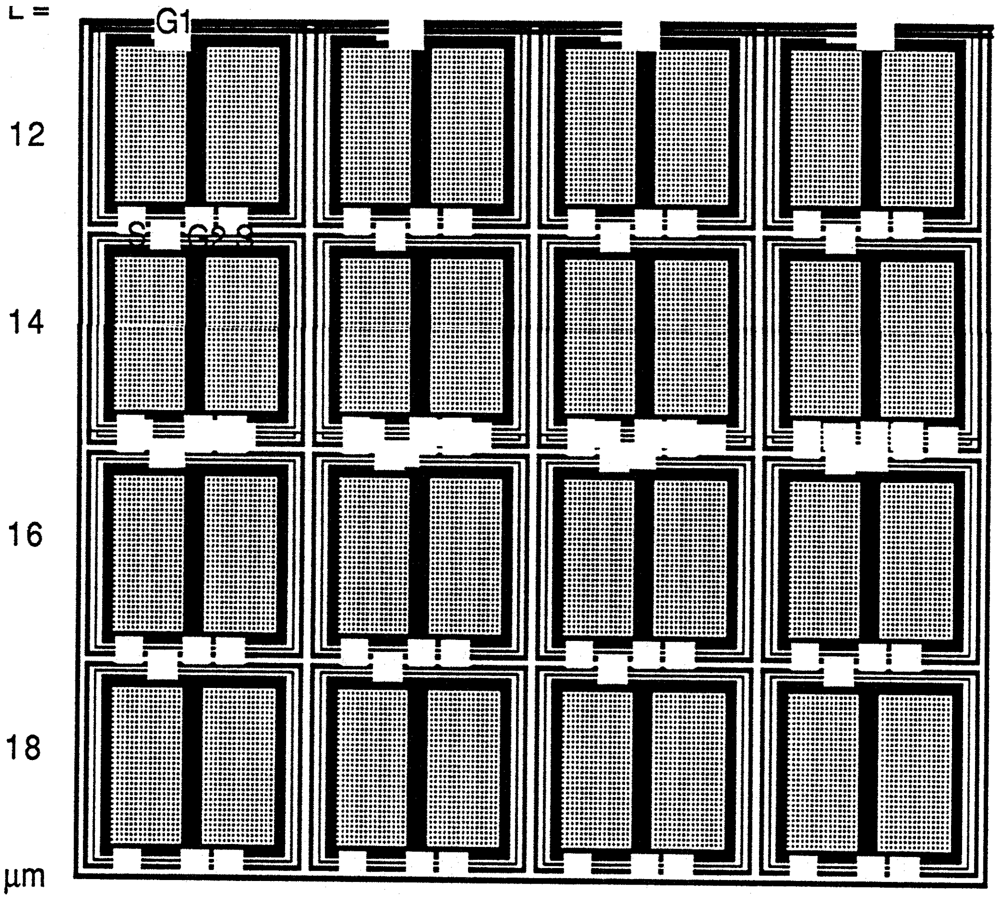


Fig. 3-9 Matrix of devices used for determining the optimal geometry.

3.2. Integrating the turn-off driver

Extending the p-well beyond a portion of the active device area provides a junction isolated region in which an independent, low voltage, lateral MOSFET may be placed. This secondary device can be made with the same diffusions as the main high voltage device, as long as the gate of the shorting switch is narrow enough for the lateral p-well diffusions to join from either end. Fig. 3-10 illustrates how the driver is incorporated within the main device with 4 µm wide gates on either side of the n+ sources. The isolated n+ diffusion between the two gates (G2) acts as the drain of Q2 and must be electrically connected to G1. The lateral spread of the n+ diffusion under the gate is 1 µm, while that of the p-well is 3 µm, giving a 2 µm channel length for the device which will be used as the integrated turn-off driver.

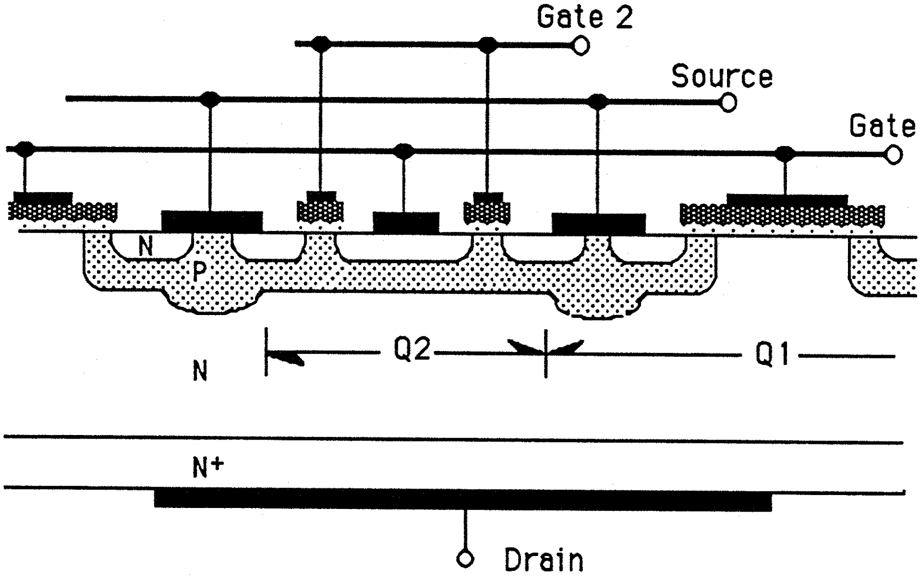


Fig. 3-10 Cross section of the integrated driver, housed in an extended p-well within the basic DMOS structure.

This arrangement allows the driver to be routed around the chip in order to achieve sufficient channel width (Z_{ch}). The lateral MOSFET must occupy as much area as needed for its full ON resistance, plus the resistance of the main device's gate poly, to meet the criterion of eq. 2-5 (4Ω for the device of this study).

For the designed device, the voltage at the gate G1 must be able to withstand 20 volts without breaking down or punching through. Since the channel length is 2 µm, the p-type doping must be 10^{17} cm^{-3} at the surface of the well.[1] If the doping is much less than that, the drain of Q2 will reach through to the source, and the gate voltage will not be able to go higher without current flowing to the source. Similarly, if the doping were made much heavier, the p-n junction would break down due to normal avalanching.

The thresholdvoltage of both the shorting switch and the main device will be the same since they are formed by the same diffusions and have equal surface doping concentration and gate oxide thickness. Assuming this concentration is approximately 10^{17} cm^{-3} and the oxide thickness is 1000Å, the threshold voltage V_t should be around 4.5 volts. For voltages greater than this, inversion charge accu - mulates at the silicon surface, and the total charge in the channel (Q_{ch}) is approximately

$$Q_{ch} = C_{ox} [V_G - V_t] \qquad [3\text{-}1]$$

where C_{ox} is the gate oxide capacitance per unit area. Knowing the inversion charge for a given applied gate voltage and the channel mobility (μ_{ch}), the resistivity in Ω/' is

$$R_{ch} = (\mu_{ch} Q_{ch})^{-1} \quad . \qquad [3\text{-}2]$$

The inversion charge is directly proportional to the applied voltage, and therefore the E-field, but the mobility is not. The dependence of the channel mobility on transverse E-field (E_x) is approximately linear for E_x between 2 and 8 x 10^5 V/cm. Experimentally determined values for μ_{ch} verses E_x at room temperature, within these limits, can be fit into the following relation,

$$\mu_{ch} = 700 - 5.56 \times 10^{-4} E_x \qquad [3\text{-}3]$$

where $$E_x = (V_G - V_t) / T_{ox} \quad .$$

A quadratic relation results in which R_{ch} is minimized when $E_x = 6.3 \times 10^5$ V/cm. Raising the gate voltage above this point only decreases the channel conductivity. Consequently, the gate voltage which minimizes the channel resistance is approximately 10 V, resulting in a minimum channel resistance of

$$R_{ch}\,,\text{min} = 4500\ \Omega/\square \quad . \qquad [3\text{-}4]$$

The resistance of the poly is the largest portion of the total resistance from the gate G1 to ground. This total resistance must be 4 Ω or less for the designed device. The area of the whole device is 1mm x 1mm to achieve the desired 0.5 Ω ON resistance from the drain to source of Q1, as required by the circuit of the case study (Ref. Chapter 2). Assuming 12 Ω/□polysilicon, and a minimum distance between metal contacts of 0.5 mm, the total gate poly resistance can be approximated for the 1mm x 1mm device.

By assuming that half the area is insulated, the effective sheet resistance is twice that of a sheet of poly. The number of squares can be conservatively approximated as half the width of the gate area divided by the perimeter. Since the distance to the center is 0.25 mm and the perimeter is 6 mm, there are 1/24 squares. Therefore, the resistance is 1 Ω by a conservative estimate, so to achieve a minimum ON resistance of less than 2 Ω, the total channel width Z_{ch} must be larger than 0.9 cm.

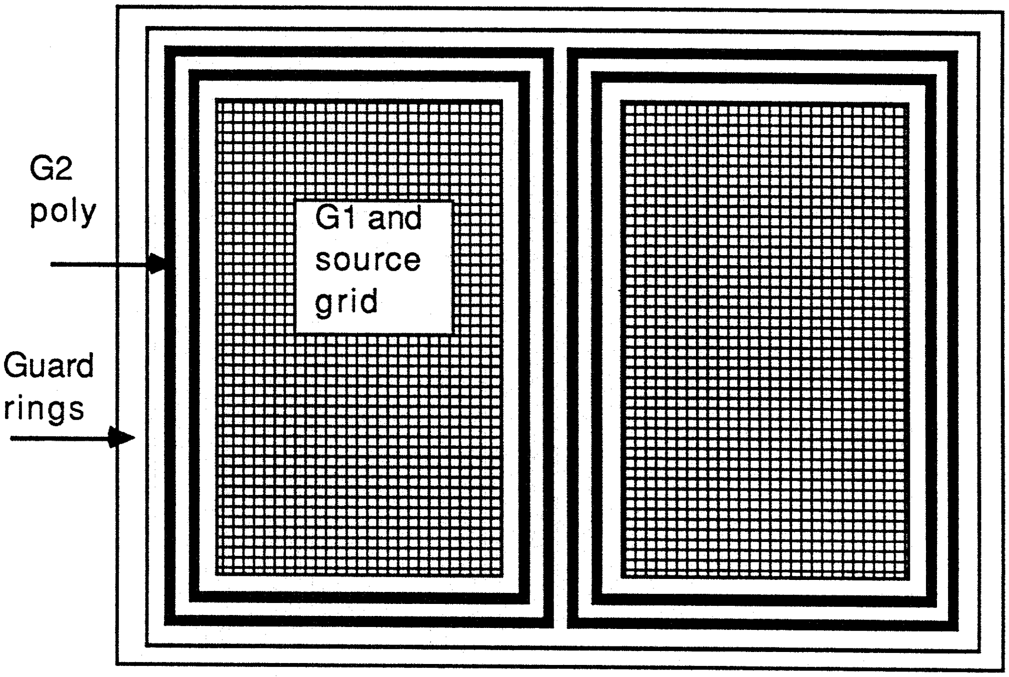


Fig. 3-11 Basic scheme for incorporating the driver around two blocks of DMOS.

The chosen configuration, as illustrated in Fig. 3-11, has two blocks of DMOS surrounded by double shorting switches. Each block is 1mm x 0.5mm, and the extended p-wells contain two gate lengths. The center of the two blocks has a row of four G2 lines separating them. The total effective channel width is therefore

$$Z_{ch} = 1.2 \text{ cm} \quad .$$

Since the channel length is 2 µm, the minimum ON resistance with 10 volts at G2 is

$$R_{ON}, Q2 = 0.75 \, \Omega$$

which makes the total gate resistance less than 2 Ω from the gate to ground when the driver is ON, as long as the connections with the metal make contact with a low enough resistance.

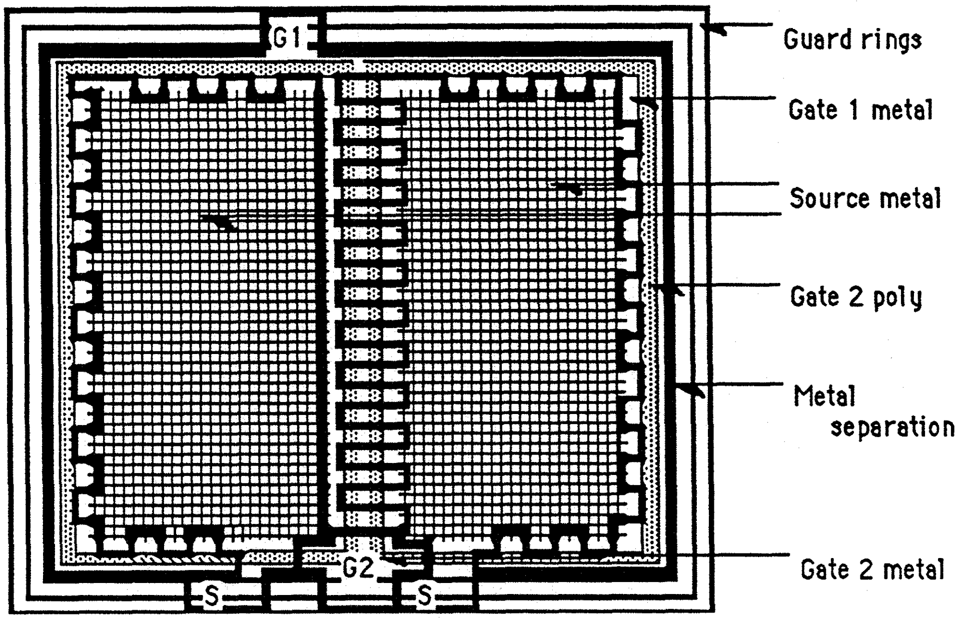


Fig. 3-12 A detail of a single DMOS with an integrated shorting switch showing the interdigitated metalization scheme connecting the source and gate to the driver.

In order to achieve a minimal resistance from G1 to the shorting switch, the metal contacts must be interdigitated as closely as possible. Fig. 3-12 illustrates how the metal connects over the shorting switch. A close-up of the center portion is shown in Fig. 3-13. The contacts are made through the field isolation oxide. Long fingers down the middle of the device contact the poly gates on either side to each other and to the n+ diffusion in the center of the shorting switch (drain of Q2).

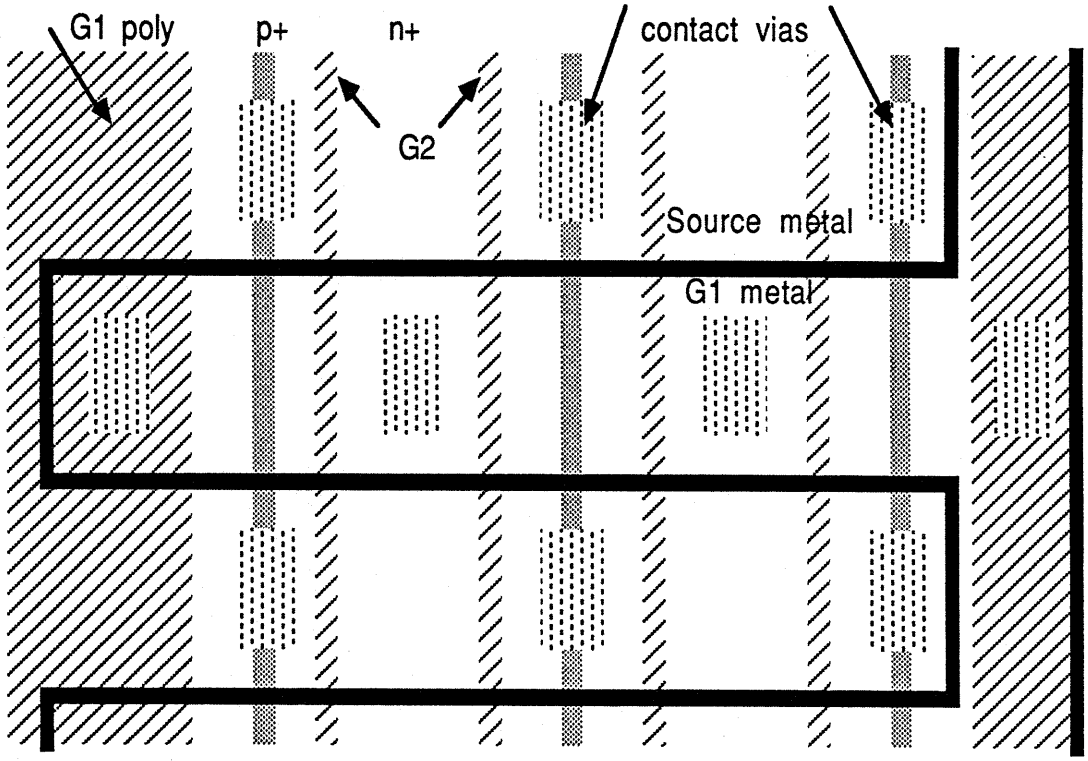


Fig. 3-13 Metalization detail in the center of the device. This illustrates how the double shorting switch is incorporated.

3.3. Calculated properties of the design

To approximate the resistances and the capacitances of the total structure, certain assumptions must be made about the device. First, each device in the 4 x 4 matrix is assumed to have only slightly different dimensions. Second, the actual values of the on-state resistance cannot be accurately modeled, so for the purpose of determining the approximate properties of the designed devices, a single case example will be used, one which should be optimized for high conductiviy per given area. The dimensions for this case example are L=16 µm, and W=8 µm. The single device contains N=1600 source wells in the 1mm x 1mm area.

The on-state resistance of the main device can be calculated by determining the conductance per source cell, multiplying it by N, and then inverting the result. The length of the p-well on one side is the opening (W) plus the lateral diffusion on either side (6μm). This makes the total p-well width (l_s) equal to 14 μm, for this example, and the separation between the p-wells (l_c) equal to 10 μm. Referring to Fig. 3-14, the total resistance includes the channel resistance R_{ch}, the resistance due to the lateral current flow in the accumulated n-region R_1, the resistance due to the vertical path between p-wells R_2, and finally, the resistance due to the spreading under the p-wells R_3.[4]

The channel has the same sheet resistance as the shorting switch: 4500 Ω/°. The total channel resistance is therefore 0.1 Ω. R_1 and R_2 may be combined, and approximated as a single resistance contribution[5,6] whose resistance equals approximately 0.1 Ω. Finally, the spreading resistance is calculated assuming a 45 ° spreading angle beneath the p-well regions. Assuming the gate over the n-region encompasses 65% of the total device area, the relative resistance has been calculated as a ratio of $R/R_o = 1.2$, where R_o is simply the resistance of the bulk epi.[7] The epi is 2 Ω-cm material, and there is a 10 μm separation between the p-well junction and the n+ substrate, making $R_3 = 0.24$ Ω. Thus, the total on-state resistance for the device is just under 0.5 Ω.

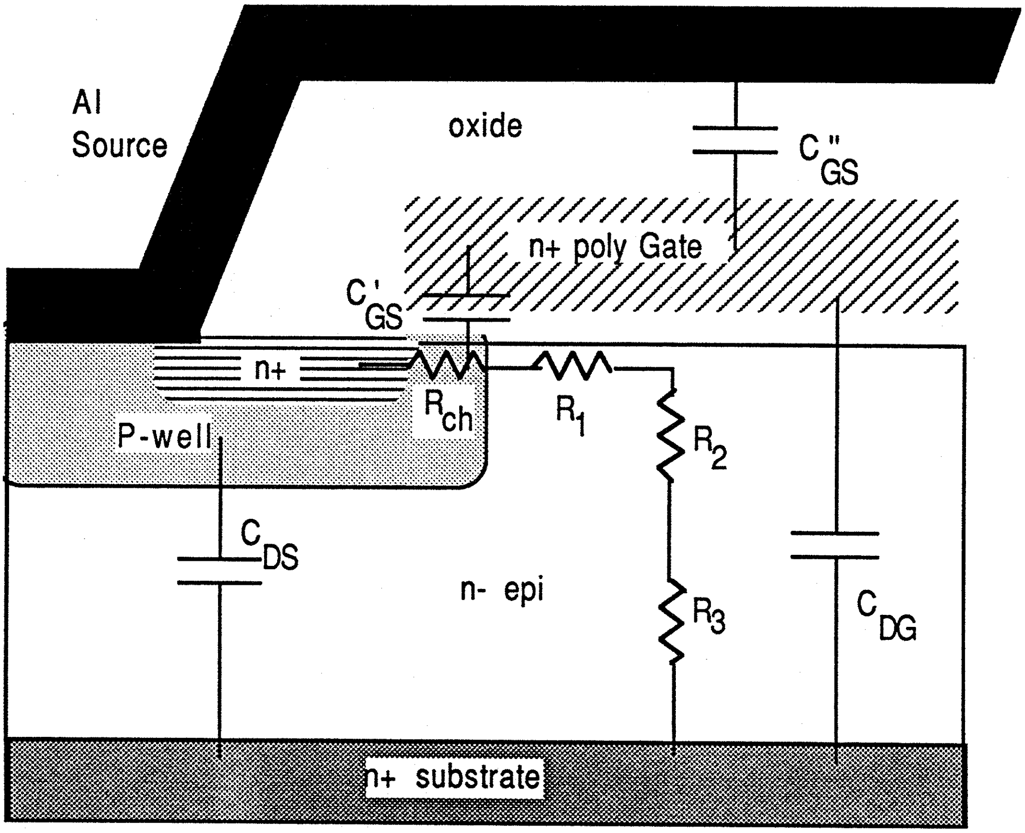


Fig. 3-14 Cross-section of the vertical DMOS showing the resistances and the capacitances for the main structure

The gate area is what determines the relative input capacitances of the two MOSFETs. These are found for the 1mm x 1mm device, where area of the main device is

$$A_{G1} = (1\text{mm})^2 - N(W)^2 = 9.0 \times 10^{-3} \text{ cm}^2 \qquad [3\text{-}5]$$

similarly, the area of the shorting switch gate (G2) is

$$A_{G2} = (12\text{mm}) \times (4\mu\text{m}) = 0.48 \times 10^{-3} \text{ cm}^2 \qquad [3\text{-}6]$$

which is simply the total length times the width of the driver. The input capacitance to the driver is thus approximately 5% that of the main device. This is important when considering the speed at which the device can turn on and off. The RC time constant of the driver, which determines the minimum time needed to create the channel, is

$$t_{RC} = C_{ox} R_{s,poly} (l_{G2}/8)^2 \qquad [3\text{-}7]$$

where R_s is the sheet resistance of the poly, and l_{G2} is the length of G2. For the designed device, this product is approximately 2 ns.

The drain - source and the drain - gate capacitances will now be calculated. They are both strongly dependent on the device area and not on the source cell geometry. Their values are significantly larger when the device is ON and the depletion region is narrow than when the device is OFF. Both capacitances follow a $V^{-1/2}$ dependence with the Drain voltage since the depletion width primarily determines the total capacitance. As described in Chapter 2, the relative values of C_{DS} and C_{DG} (the miller capacitor) are nearly constant by a factor ç. This geometrical factor relates the area beneath the p-well to the surface beneath the gate in the n-layer. The results for the device are summarized for each steady state (ON and OFF):

State	V_{DS}	C_{DS}	C_{DG}
ON	1 volt	100 pf	240 pf
OFF	150 volt	3.4 pf	6.6 pf

Finally, the gate - source capacitance calculation considers the driver's contribution to the total capacitance. Fig. 3-15 illustrates the device cross section including the driver. The contribution to the gate - source capacitance from the shorting switch is shown schematically as C_{GS}'''. This capacitance depends strongly on the gate (G1) - source voltage and must be added to the contributions from the isolation oxide and the gate overlap, as shown in Fig. 3-14 (C_{GS}' and C_{GS}'').

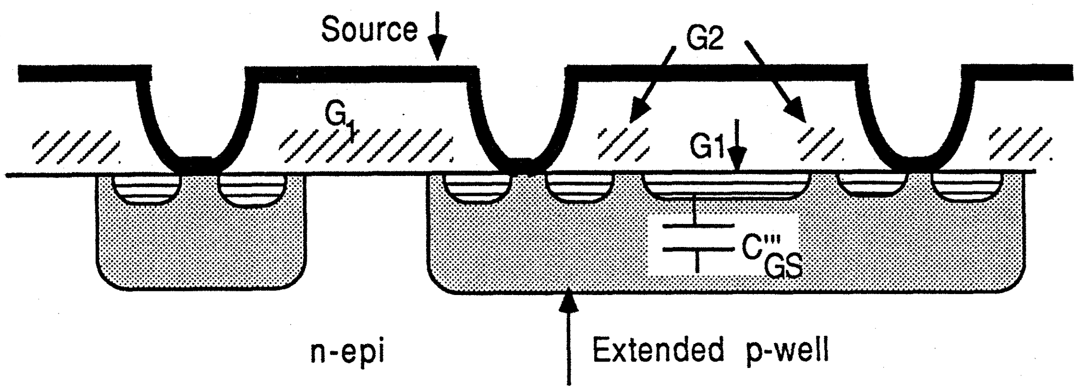


Fig. 3-15 cross section showing the final metalizatin, indicating the depletion capacitance contribution form the driver.

The isolation oxide thickness is approximately 1.5 μm thick over the entire structure. Assuming the area of this capacitor as A_{G1} gives a value of $C_{GS}' = 21$ pf. Similarly, the fixed capacitor due to the the overlap of the gate poly with the p-well diffusion yield a value $C_{GS}'' = 66$ pf. A more rigourous calculation is not necessary since these capacitances are nearly independent of voltage, and they are generally small by comparison to the depletion capacitance of the shorting switch.

Depletion in the shorting switch only occurs beneath the drain of Q2 since the source is shorted to the p-well, so the area of concern is the junction of the n+ diffusion with the p-well. The calculation assumes a uniform doping of 10^{17} cm^{-3} with an area of 12 μm wide and 12 mm long and a uniform, one-dimensional depletion layer. When the main device is ON, the gate voltage (G1) is high at 10 V and this capacitance is found to be 39 pf. The value increases with $(V_{G1} - 1)^{-1/2}$ to its maximum capacitance of 130 pf. Thus, summing the three components of the Gate to Source capacitance together yields:

State	V_{G1}	G_{GS}
ON	10 volts	126 pf
OFF	0 volts	215 pf

CHAPTER IV.

THE FABRICATION PROCESS

Each high temperature step in the fabrication process increases the likelihood that impurities will enter the wafer. In order to minimize extraneous contamination, the process sequence uses the fewest possible number of diffusion steps by combining the drive-ins of the n+ source and the p+ contact into the same step. This approach allows the p+ plug, which is self-aligned to the source contact, to be diffused only a small amount. It is only driven in with the final diffusion step, so it does not spread any farther than about 2.5 μm under the source diffusion. The n+ region is also more heavily doped than the p+ region so the metal will make a good contact to the source and the p-well. The source and p-well can thus be made much smaller than is typical for devices where the p+ contact is driven in first.

Since the p-wells and sources are self-aligned to the gate, the gate oxide growth and the polysilicon deposition must take place before any of the dopants are implanted and diffused. The fabrication thus begins with lightly doped n-type silicon epi wafers on (100) n+ silicon substrates. The epi is of the thickness and resistivity needed for standing off the 150 V required. Aside from the epitaxy, high temperature steps include: oxidation, polysilicon deposition, p-well diffusion, n+ and p+ diffusion, and deposited oxide densification. Other steps, including oxide deposition, are performed at temperatures below 500 °C where they do not contribute significantly to dopant diffusion or wafer contamination.

The following five sections of this chapter discuss details of the fabrication sequence. The appendix contains a complete step-by-step proceedure for fabricating the actual devices whereas this chapter discusses the purpose of each step in the overall process.

First, the gate oxide is grown and polysilicon is deposited. The poly is then patterned and plasma etched to form the basic structures. Boron is implanted through the openings and then diffused to nearly its final depth in the silicon. Through the same openings, Arsenic is implanted but not driven in. The poly is also doped with the arsenic by the same implantation.

The poly surrounding the active device area, and between the field guard rings, is then removed by a second plasma etch, through openings defined by the second mask. At this point, the poly is doped, so it has a different etch rate than pure poly, hence the etch rate must be monitored before each run.

A 1 μm layer of oxide is then deposited over the structure and patterned with the third mask. Openings to the contact plugs, through the oxide, help align the final poly etch. Once the oxide over the plugs and the poly have been etched, the oxide over the rest of the wafer protects the poly from the deep Boron implant meant to be the p+ contact diffusion. After this implant, the oxide is removed so a final isolation oxide deposition will be uniform over the wafer. The oxide is then densified with the long diffusion at a low temperature (1000 °C) in a nitrogen ambient.

The wafers are then completed by defining and etching the contact vias with the fourth mask followed by the metalization with aluminum. The metal is then patterned and etched using the fifth mask and the backside of the wafer is prepared and metalized.

4.1. Gate oxide and polysilicon

The growth of 1000 Å of gate oxide on a prepared and cleaned epi wafer as the first process step insures that no sodium or other bulk impurities will be introduced into the oxide from a previous diffusion. Immediately following the oxidation, gate polysilicon is deposited by low pressure chemical vapor deposition (LPCVD). This passivates the oxide by encapsulating it between the silicon epi and the polysilicon gate. Also, if the wafers go directly from the oxidation furnace to the poly furnace, the likelihood of impurities being introduced is further decreased.

Once the wafers have been properly prepared for oxidation, and the furnace and boat have been cleaned in an HCl environment for at least 1 hour at 1100 °C, the wafers are loaded on the boat and put into the furnace. The parameters for growing 1000 Å of gate oxide have been experimentally determined for the furnace in the microfabrication laboratory at MIT to be 46 minutes at 1100 °C in dry oxygen. This process was first performed on test wafers which were then measured to have oxide thicknesses of 1020 ± 30 Å, a sufficient value for the desired gate oxide.

During the oxide growth, the polysilicon furnace is prepared for the 1 μm deposition. When the growth is completed, the wafers are brought to the mouth of the oxidation furnace where they remain in the nitrogen flow until they can be loaded into the poly furnace. This insures that they do not become contaminated between the two steps. The wafers are then loaded onto the boat of the poly furnace, without delay, after being withdrawn from the oxidation furnace. No wafer cleaning is needed between these two steps.

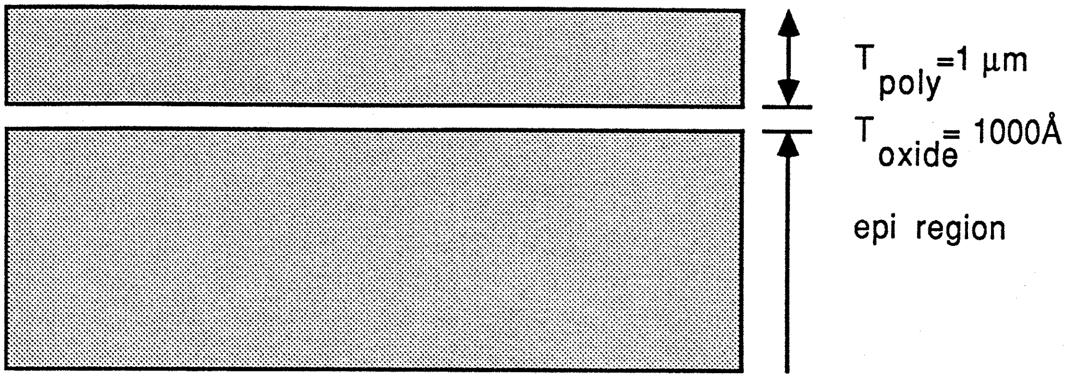


Fig. 4-1 Cross section of the device with the gate oxide and polysilicon.

The structure at this point is shown in Fig. 4-1, where the 1000 Å of gate oxide on the n- silicon epi is covered by 1 µm of polysilicon. This sandwich is patterned by the first mask and the photo-resist alone is what masks the poly for the plasma etch. The wafers are etched in a Freon plasma in the LAM plasma etching system, where silicon is selectively removed. The etch rate of polysilicon is nearly five times that of the underlying oxide, so the process is essentially self stopping at the oxide under the poly. This process step is further described in section 4.4.

4.2. Double diffusion

The p-well and the n+ source are both implanted through the same openings created by the first lithographic step. The Boron must first be diffused in deeply enough so there will be sufficient isolation between the drain and the source after the second diffusion (Fig. 4-2). The second diffusion must drive in the p+ contact region as deeply as possible while not allowing the n+ source to diffuse lateraly more than 1 µm. Since the diffusion coefficient for Arsenic is approximately 5 times smaller than that of Boron at 1000 °C, the second diffusion step will be done at this temperature for long enough time to make the junction between the n+ source and the p-well occur 1 µm laterally under the gate.

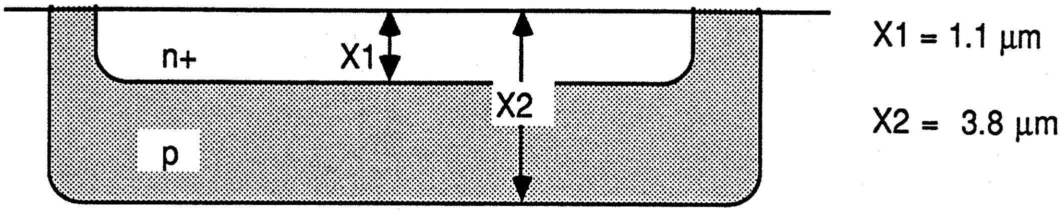


Fig. 4-2 Cross section of the source and p-well diffusions.

For a 2 µm channel, the p-well diffusion extends 3 µm laterally under the gate and has a doping of 10^{17} at a point 1 µm into the region where the junction forms with the source. According to Colclasser[3], with a junction concentration of 10^{15}, the lateral junction is 0.78 that of the vertical junction. Similarly, for the source diffusion, with a junction doping of 10^{17} and a surface concentration of 10^{21}, the lateral junction depth is 0.9 of the vertical depth. This gives us a desired profile as shown in Fig. 4-2 with the p to n- epi junction at 3.8 µm deep, and the n+ to p-well junction depth at 1.1 µm.

An initial SUPREM simulation suggested that the following sequence would yield the desired profile: a Boron dose of 3×10^{13} cm^{-2} with an energy of 50 keV, followed by an initial drive in at 1150 °C for 5 hours, then an Arsenic implant dose of 1×10^{16} cm^{-2} at 150 keV followed by a drive in at 1000 °C for 20 hours. The results of this simulation are shown in Fig. 4-3. However, when this sequence was performed on actual monitor wafers, the n+ diffusion was found to occur at 1.2 µm, and the p diffusion was at 3.5 µm.

After several iterations, a final sequence was developed which produced the desired profile of Figs. 4-2 and 4-3. This includes a Boron implant dose of 5×10^{13} cm^{-2} with an energy of 50 keV and a 6 hour drive in at 1150 °C, followed by an Arsenic dose of 1×10^{16} cm^{-2} at 150 keV and a 15 hour drive in at 1000 °C.

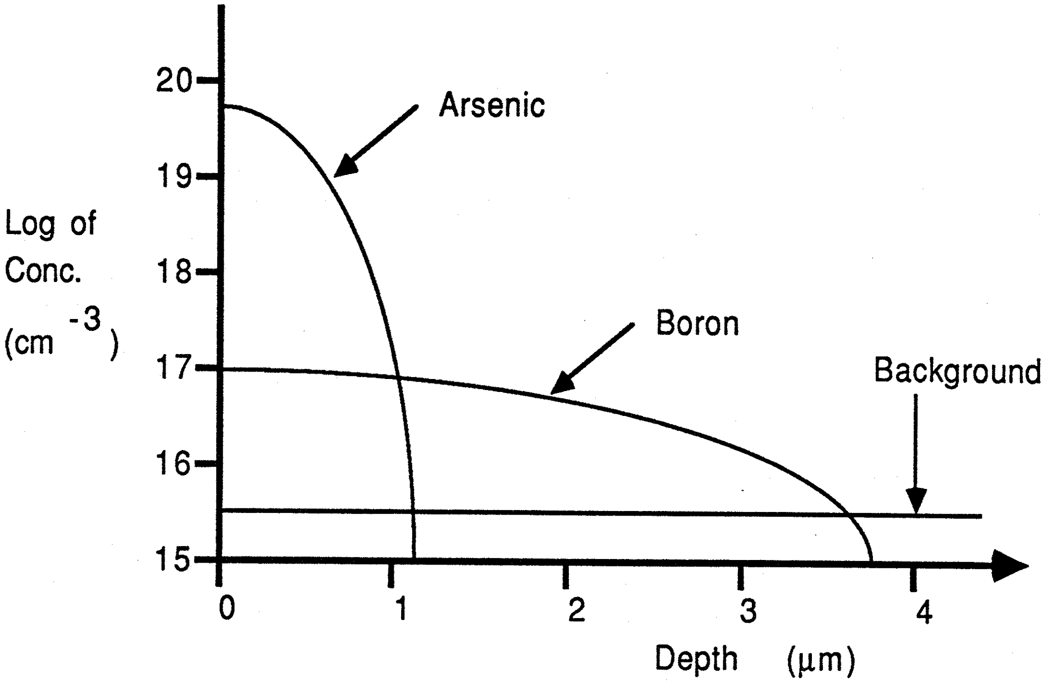


Fig. 4-3 SUPREM simulation for the required doping profile.

4.3. P+ source contact

The heavily doped p+ contact must diffuse deeply into the p-well with only the final drive-in step. The highest possible implant energy available for the 2" wafers used is 200 keV which has a projected range for Boron of 0.7 µm. A dose of 5 x 10^{15} cm^{-2} at that maximum energy should diffuse approximately 2.5 µm into the silicon, resulting in the profile shown in Fig. 4-4. This was found to be very near the actual profile produced in monitor wafers.

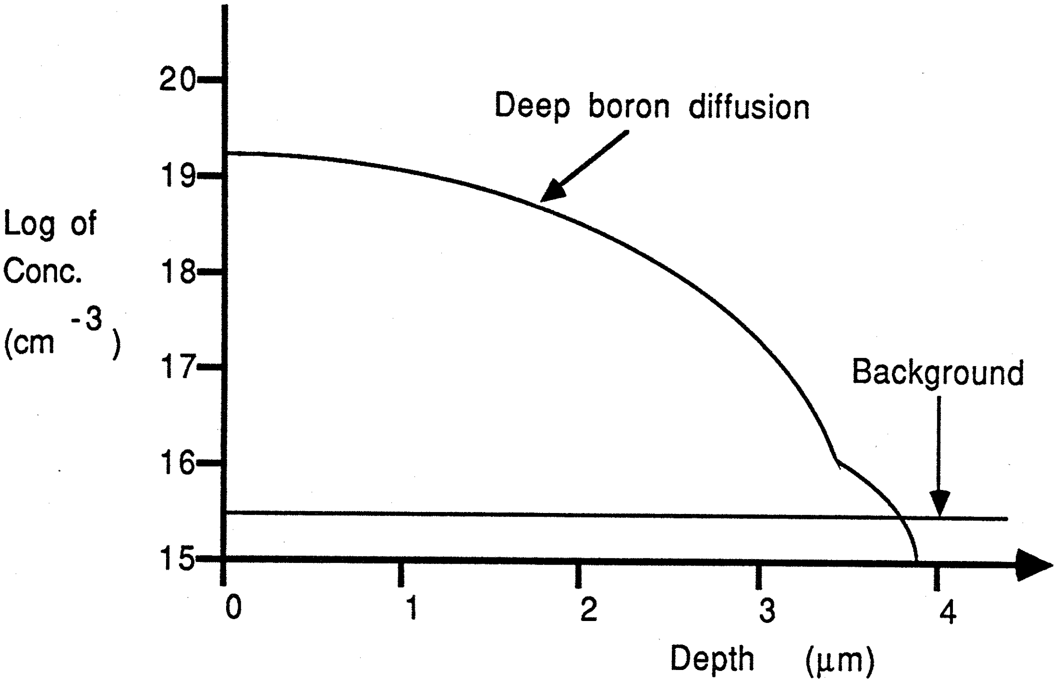


Fig. 4-4 SUPREM simulation of the doping profile under the p+ contact.

4.4. Polysilicon etch

The LAM plasma etching system in the microfabrication laboratory at MIT is anisotropic and has a selectivity of undoped silicon to oxide of approximately 5 : 1. This allows one to reproduce a pattern from the resist onto the poly and have the etch stop at the oxide (Fig. 4-5). An end-point detection system checks for oxygen in the plasma, indicating when the etch has penetrated the poly. Stopping the etch when oxygen is detected by the end-point probe prevents the oxide from being etched. However, the etch rate is not completely uniform, so the plasma is only stopped when the oxygen content reaches a constant level. Then, a short over-etch insures that the poly has been removed.

The etch rate of doped poly is higher than that of undoped poly, thus, two different etch times are needed for the different points of the process. The first mask level identifies the double diffusion and the openings for the p+ implant where 1 μm of undoped poly becomes etched. Since the poly is doped with the implantation of the n+ source, the second two poly removal steps must use the etching times for 1 μm of heavily doped n+ poly.

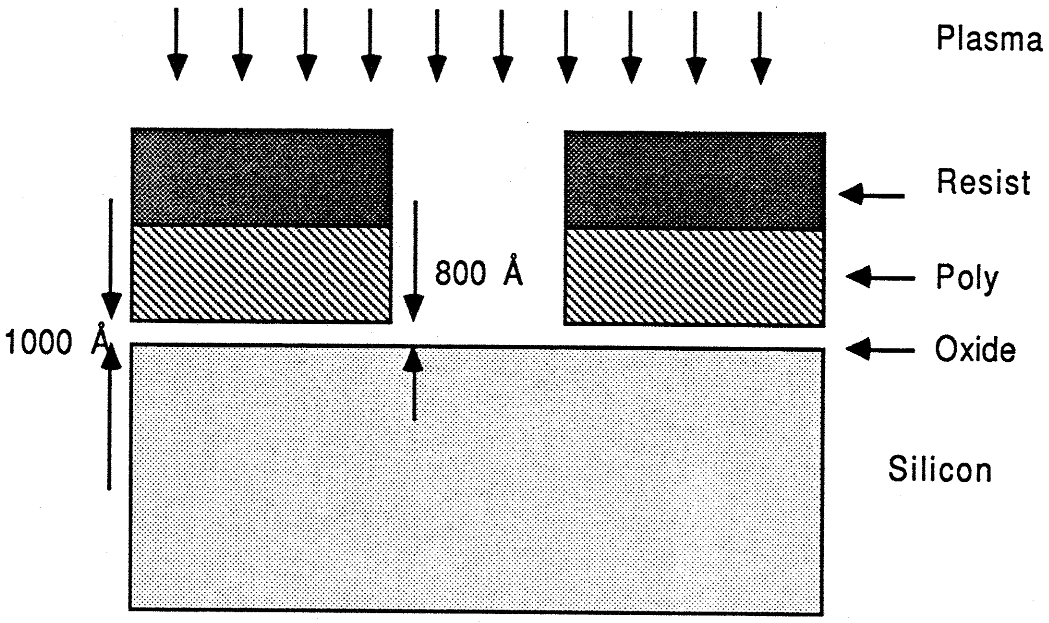


Fig. 4-5 Cross section of etched polysilicon showing the selectivity and anisotropy.

The etch times were measured from unpatterned wafers with 1000 Å of oxide and 1 μm of poly. The end-point detection system determined when the etches had been completed and they were followed by a 10% overetch. The etching times for the poly were found to be 4 minutes for the undoped samples, and 3:15 for the doped samples. The remaining oxide was found to be approximately 800 Å thick.

4.5. Isolation oxide etch

1.5 μm of CVD oxide is deposited with the PYROX oxide deposition system for the isolation oxide. However, the oxide cannot be reliably deposited on bare silicon in a layer of more than 1 μm thick. Consequently, the oxide must be deposited in two steps. First, 0.5 μm is deposited and densified at 900 °C for 30 minutes in an oxygen environment. Then, a second layer of 1 μm is deposited and densified in the same manner.

When oxide is chemically etched on silicon, one determines its completion by observing when the acid (Buffered HF) sheaths off the wafer. However, a wet etch is isotropic and etches laterally under the resist as well as vertically.

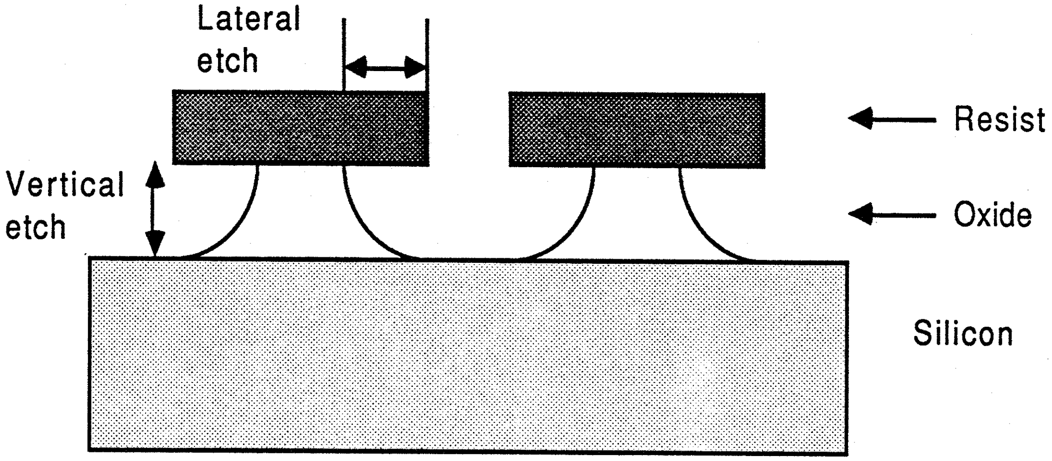


Fig. 4-6 Cross section of patterned and etched isolation oxide on a bear wafer.

The degree of isotropy is determined by measuring patterned lines on deposited oxide before the wafer is etched and the resulting pattern afterwards. The difference is the amount of lateral oxide etched (Fig. 4-6) which is divided by the oxide thickness to get a ratio of the lateral etch rate to the vertical etch rate. This ratio has been determined to be approximately 1 for both densified and undensified oxide. The etch rate, as determined by the time needed to etch the oxide, was found to be approximately 1800 Å/minute for densified CVD oxide, and 1000 Å/minute for thermal oxide at room temperature.

For the devices of this study, a lateral etch of 1.5 μm is not tolerable in the smallest of the source cells (W=6 in Fig. 3-3). There is only 2 μm between the p+

contact and the edge of the poly gate, leaving 0.25 μm on each side for alignment tolerance. This pushes the limits of the CGA wafer stepper, and is not suitable for fabrication.

The lateral etch, however, can be nearly halved by using a double etch scheme. The first time the resist is patterned by the fourth mask, half the oxide is etched, so 0.75 μm of oxide is etched laterally as well as vertically. Photo-resist is then reapplied and the mask pattern is exposed once again (Fig. 4-7). Finally, the rest of the oxide is removed beneath the second application of the same mask.

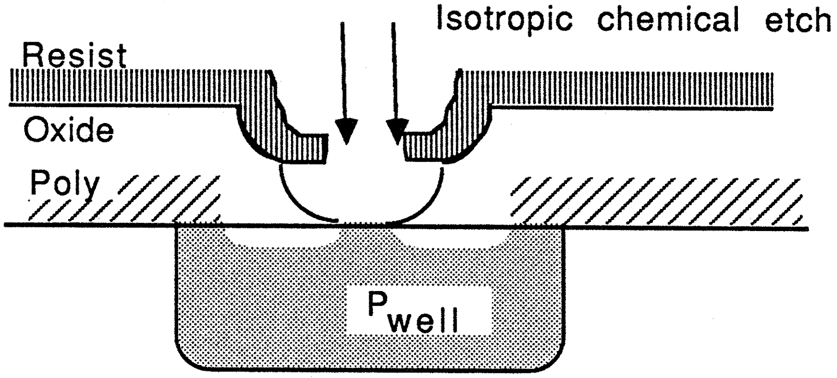


Fig. 4-7 Cross section of the double wet etch method employed by the process

Such a scheme allows one to use tighter design tolerances for the metal contacts to the silicon than if a simple, single etch scheme were employed. This technique prevents the source from shorting with the gate by nearly halving the lateral etch and making it possible to have cell widths (W) as small as 6 μm.

CHAPTER V.

MASK DETAILS

Tolerances must be incorporated into the design of a set of masks which are aligned to one another. The smallest features must be able to accommodate the largest alignment error, which then limits the minimum dimensions of the structure. However, self-alignment techniques allow one to relax the necessary tolerances so the overall dimensions can be made smaller.

The fabrication process of the vertical DMOS described in this study uses source diffusions which are self-aligned to the gate and p+ contacts which are aligned to the centers of the source regions. This technique assures that the mask dimensions are limited by size of the contacts to the source diffusions and by the thickness of the isolation oxide, rather than by the alignment tolerances to the p+ contact.

The process accommodates a p-well (W, as shown in Fig. 3-3) whose minimum dimension is determined by the lateral etch of the final isolation oxide and its alignment to the rest of the structure. This oxide is approximately 1.5 μm thick, so with a perfectly isotropic etch, the minimum spacing between the source contact and the gate is over 2 μm. The double etch technique described in section 4.5, however, allows the lateral etch to be less than 1 μm on each side. Combined with the given 0.25 μm alignment tolerance of the GCA wafer stepper for each mask layer, a minimum acceptable space outside of the "plug" region (labeled P in Fig. 3-3), is 2 μm on each side. Therefore, the smallest dimension for W is 6 μm; 2 μm for P and a 2 μm separation from the gate poly on each side.

Starting with the smallest value of W=6 μm with P=2 μm, the array of Ws making up the matrix of Fig. 3-5 gets larger by increments of 2 μm, up to 12 μm. This is accomplished by increasing the spacing between the plug and the gate to 3 μm for W=8, then by increasing the plug to 4μm for W=8μm, and finally by increasing the spacing to 4 μm with the 4 μm plug resulting in W=12 μm. This largest dimension should certainly be resolved with the equipment in the microfabrication facility at MIT, and is used for the shorting switches on all the devices.

5.1. Oversized flashes

Since positive photoresist was to be used for all the layers, the mask set was initially designed using clear field projection reticals on the GCA wafer stepper. The reticals are at 5x the actual size, since the stepper reduces the image by that amount during the projection of the pattern on the wafer. The masks were made on the CAD system by identifying flashes (exposures) everywhere resist was to be left on the chip.

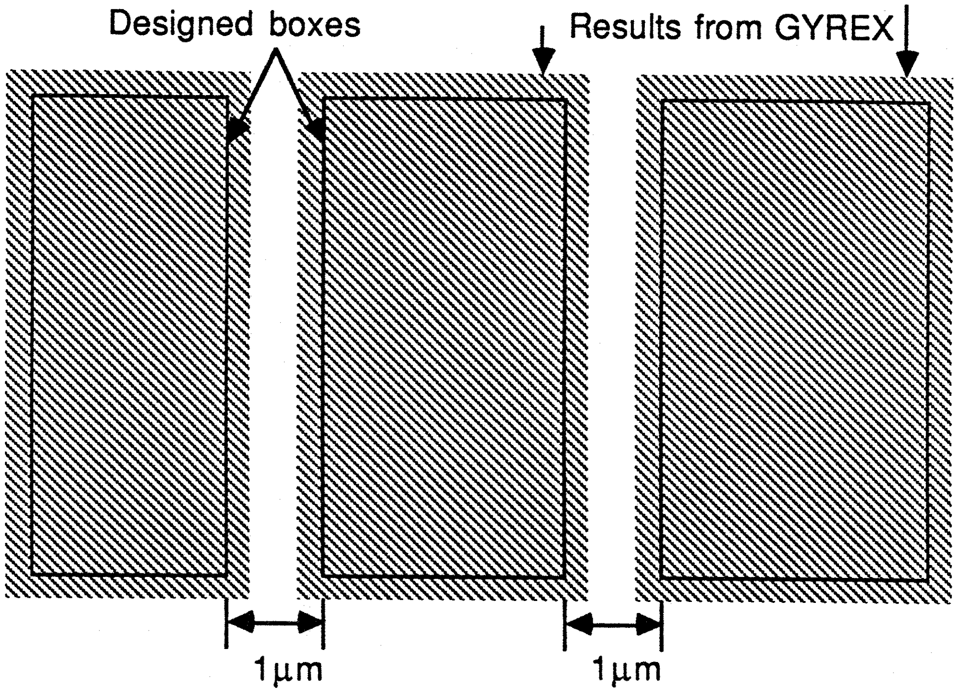


Fig. 5-1 Illustration of how oversized flashes narrowed the guard ring openings.

The GYREX mask exposure equipment, used to produce the reticals, oversizes the flashes by 1µm in each dimension. Oversizing insures that adjacent flashes butt end to end by allowing them to overlap slightly. However, this makes intentional spaces between flashes narrower than may be desired. For example, the guard rings were formed by flashes which were supposed to be separated by 5 µm on the retical (1 µm on the chip). The flashes were oversized by 1µm on either side, making the separation only 3µm. When reduced on the wafer, this became 0.6 µm (Fig. 5-1) which was not well resolved around the entire chip.

Because of the over-sized flashes, the smallest openings on a clear field mask must not be designed to be less than 2 µm. A 2 µm space should be 10 µm on the 5x retical, but it is reduced by 1 µm on either side from the over-sized flashes, and it actually becomes 8 µm. When this is projected on the wafer the resulting space is only 1.6 µm wide. The width of the guard rings, as a consequence, is limited by the minimum practical spacing between flashes, and should be designed with a minimum of a 2μm opening around the device area.

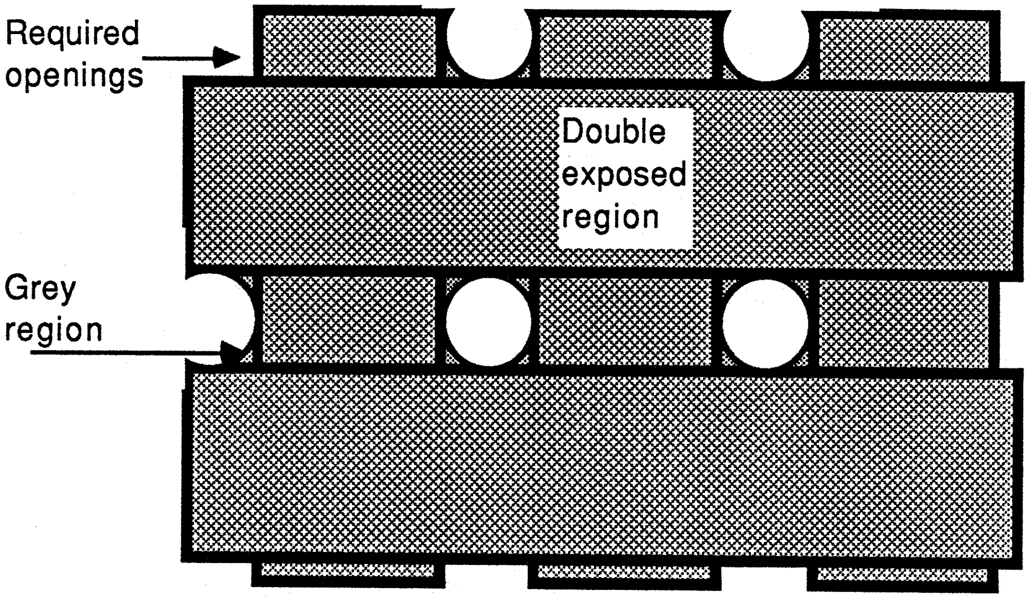


Fig. 5-2 overlapping flashes used to define openings on the clear field reticals.

The minimum dimensions for the third and fourth mask layers, which indicate the openings to the center of the source regions, were made with overlapping vertical and horizontal flashes as shown in Fig. 5-2. The spacing between adjacent boxes is 2 μm, and should be resolved as 1.6 μm openings, as explained above. However, a second effect results which further degrades the desired openings. The emulsion on the retical is double exposed everywhere two lines overlap. The double exposed region tends to leak into the spaces, exposing the emulsion. This makes the edges of the openings on the retical grey, so the actual openings are smaller than desired, and the 2 μm spaces were nearly unresolved.

The resolution of these openings can be improved by using dark field reticals with the small boxes in the center to define the openings. The same GYREX retical exposure unit oversizes the flashes which produce the spaces within a dark field.

Such a technique assures that there will be no double exposures. The clear holes in the center will be slightly oversized which is desirable for both layers 3 and 4. Also, since the metal is defined by exposing the resist where the lines are to be separated, layer 5 should also be made with a dark field retical.

When designing the dark field masks, it is important to realize that where boxes are drawn on the CAD tools, openings will appear on the mask, hence openings will be made in the resist once stepped on the wafer. Therefore, the mask which defines the poly (layer #1) uses a clear field retical, while the layers which define the finer source structures (layers #3 and #4) and the layer which defines the metal (layer #5) should use dark field reticals. The second layer does not have any critical dimensions, so either can be used. However, since it mostly covers the device area, it should be made with a dark field retical, as well.

5.2. Alignment markers

The alignment markers on the first mask are the only ones needed for the entire process. All subsequent masks align to the first layer since it defines the entire device. Consequently, the later masks should provide large clear areas to make sure the poly used to define the cross is clearly visible from one layer to the next. Fig.5-2 shows the scheme for defining the mask alignment markers. Silicon is shiny, and easy to see as compared to oxide, thus the first set of markers will suffice for every subsequent alignment.

The first layer places the markers so they are spaced by 38.1 mm along the horizontal on the wafer. These markers should simply be two vertical lines, 120 μm long and 2 μm wide. The second and third mask only need to assure that a large block of resist cover the pattern since they both define regions where poly is etched.

the alignment cross will be removed.

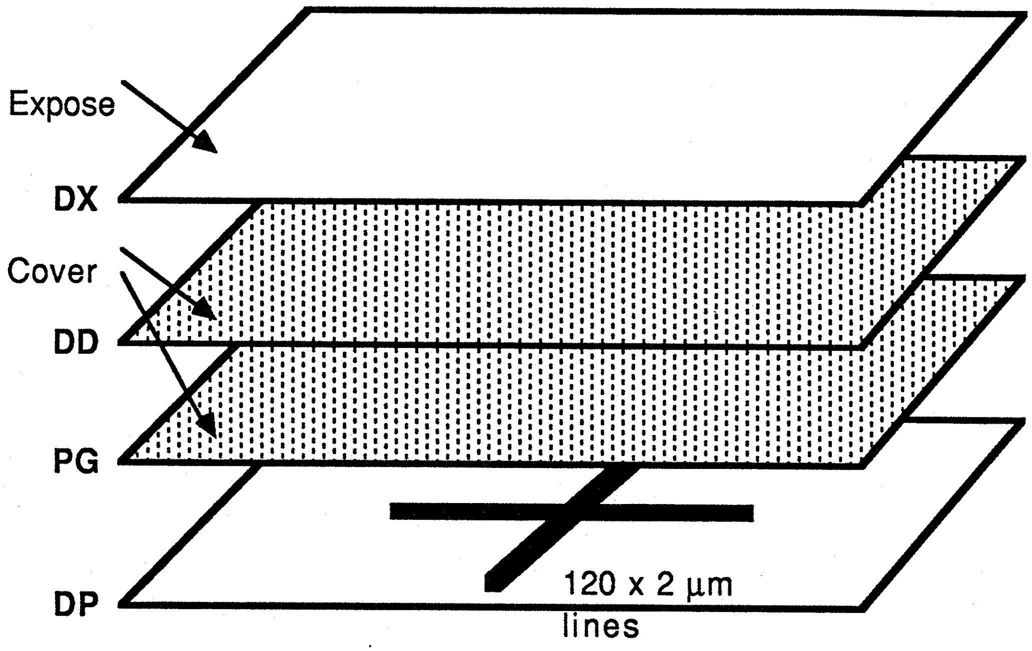


Fig. 5-3 Suggested masks over the wafer alignment mark.

This method provides a good contrast on the alignment screen which simplifies the alignment process. The key offsets for the masks will all be the same since they align to the same layer. The alignment run should still be performed before each device wafer run to insure that the baseline and the key offsets have not drifted over time. The extra time to assure proper alignment and exposure will certainly be worthwhile.

5.3. Individual mask layers

5.3.1. Mask layer #1. DP

This first mask is the most critical one since it basically defines the entire structure. It patterns the gate poly for the main device and the driver, the openings for the double source implants, and the regions for the p+ plug implant. The poly blocking the n+ from the center of the p-well is removed in a later masking step which assures the alignment of the metal contact to the center of the source region and makes contact to both the source and p-well. The alignment of the p-plug is inherent with the first mask, so the tolerance of that later mask may be large.

The guard rings are also self-aligned to the perimeter of the device since they are defined by the first mask as openings in the poly (Fig. 3-4). Boron is initially implanted through the resist on top of the etched undoped poly. The photoresist is over 1 μm thick and the implant energy is only 50 keV, so no p-type dopant enters the poly. All the Boron from the p-well implant is removed when the resist is stripped off the wafer. Then the wafer is cleaned and the Boron is driven in with the first diffusion for 6 hours at 1150 °C in a Nitrogen environment.

There should be very little oxide and no resist on the poly at this point. Consequently, the heavy arsenic dose with an energy of 150 keV becomes incorporated in the poly as well as in the source openings, doping them both simultaneously with nearly equal concentrations. The dose should approach the solid solubility limit since they will both be degenerately doped. A 1 μm thick sample (poly), implanted with a dose of 10^{16} cm^{-2}, will have an average concentration of 10^{20} cm^{-3} for both the gate poly and the n+ source.

5.3.2. Mask layer #2, PG

Once the guard rings have been defined and diffused in, it is necessary to remove the polysilicon covering the outside of the structure, between the devices. This second mask step therefore covers the entire device with resist and exposes the surrounding poly so it can be removed with a plasma etch.

The oxide through which the implants were made is approximately 800 Å thick while the poly is still 1 μm thick. Since the selectivity with the plasma is only around 5:1, the remaining oxide will be completely removed once the poly has been etched. The underlying silicon will consequently become etched and damaged. This can be avoided by covering the openings with resist from this second mask. The tolerance for this step may be as large as .5 μm on either side, since poly may remain over the guard rings, near the center, without causing any negative effects. The guard rings are at an approximate equipotential, so having the poly over that region is unaffected, as long as the poly remains over the p-diffusion. Fig. 5-4 illustrates how this mask should cover the openings in the poly which defined the field guard rings.

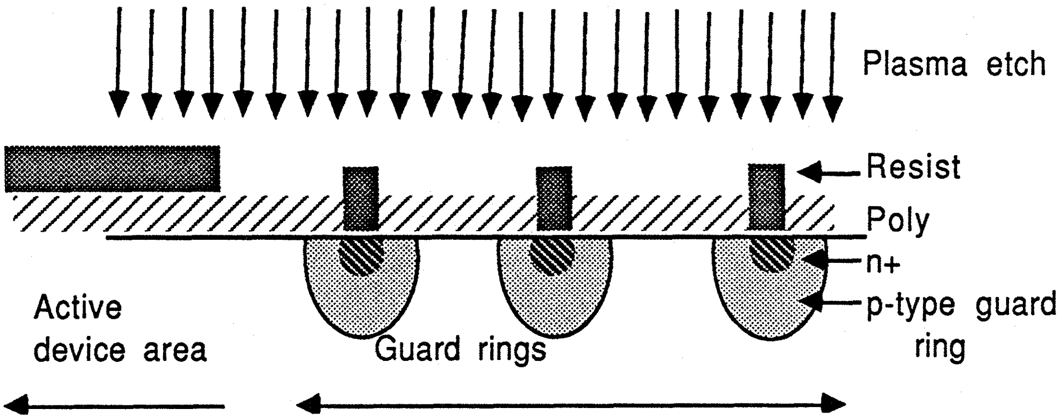


Fig. 5-4 Cross section of the guard rings which surround the device, and the resist covering the openings which had previously defined them.

5.3.3. Mask layer #3, DD

After the poly surrounding the devices has been removed, as defined by the previous mask, 1μm of oxide is deposited over the entire structure. This oxide acts to mask the poly and the silicon surface from the heavy p+ implant. It also insures the anisotropy of the poly plug removal, as defined by the third mask layer

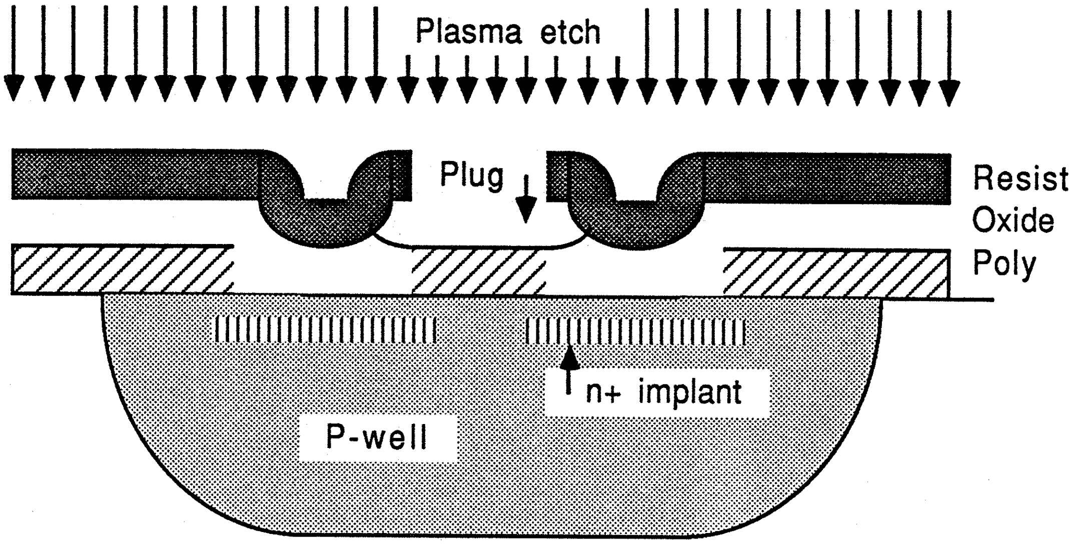


Fig. 5-5 Cross section of a single source cell just before the third poly etch step.

This deposited oxide does not need to be densified since it will be removed immediately after the boron implant. Resist covers the oxide, giving approximately 2μm of buffer from the surface to the underlying silicon. The mask defines the openings over the poly plug regions, where the oxide must first be chemically etched before the poly may be etched in the plasma. Fig. 5-5 illustrates how the mask defines the plug, after first wet etching the oxide. After the wet etch, the anisotropic plasma removes the poly down to the underlying gate oxide.

A heavy dose of boron is then implanted in the silicon, through the oxide which is left over from under the poly plug removal. The energy of this implant is 200 keV, which has a projected range of 0.5 μm into silicon.[2] The oxide under the resist assures that the implant is not introduced into the poly. Thus, after the heavy boron implant, the resist and then the undensified oxide should be etched off before the final isolation oxide is deposited to insure that excess boron does not diffuse into the poly from the oxide.

The final isolation oxide is deposited in two steps as described in section 4.5. The wafer must be thoroughly cleaned before the initial deposition, and then again before the first densification at 900 °C for 30 minutes in dry oxygen. A second layer of oxide is then deposited immediately after the wafer is removed from the oxidation furnace to avoid a second cleaning sequence. After the second half of the 1.5 µm of isolation oxide has been deposited and the wafer has been cleaned, the wafer is ready for the final long diffusion step which drives in the p+ contact and the n+ source diffusions.

5.3.4. Mask layer #4, DX

The final diffusion step also densifies the 1.5 µm of isolation oxide with a 15 hour drive-in at 1000 °C in a nitrogen environment. The oxide is nearly uniform over the source wells and the poly, so the same mask can open up the contact vias to the gates and the sources. This mask is used twice as described in section 4.5, with half the oxide being removed each time.

Once the etch reaches the silicon, the resist should be removed so the wafers can be prepared for metalization. This is done with a thorough organic and ionic clean, and with a 10 second dip in 10:1 HF to completely remove the oxide from the silicon surface. Aluminum is then deposited to a thickness of 0.75 µm with an e-beam evaporator which provides a uniform layer, directed into the contact openings.

5.3.5. Masl layer #5, DM

Finally, the metal lines are defined by the fifth layer. Separations between the two gates and the source metals are formed as shown in Fig. 3-8 and Fig. 3-9. The metal is isotropically chemical etched through openings of at least 2 µm wide. This makes the final separation between lines approximately 4 µm, due to the under-etching.

To complete the device, the backside of the wafer is etched in a freon plasma to remove the oxide and the poly, exposing the n+ substrate. Afterwards, metal must be deposited on the back and annealed at 450 °C for 30 minutes in forming gas (20% $H_{2:}$:80% N_2). This makes good ohmic contacts between the metal and the heavily doped silicon on both the drain and source contacts.

CHAPTER VI.

RESULTS

The finished wafers were tested for their static characteristics by on a curve tracer. The vacuum chuck, used to hold the wafers, served as the common drain contact to all the vertical power devices while the sources and the gates (S and G1) were connected from the top. The sources, gates and drains (S, G2, and G1) of the lateral MOSFET drivers were all accessed from the top of the wafer. This way, the power device and the drivers could be tested independently in order to determine their transfer characteristics.

6.1. Tested Devices

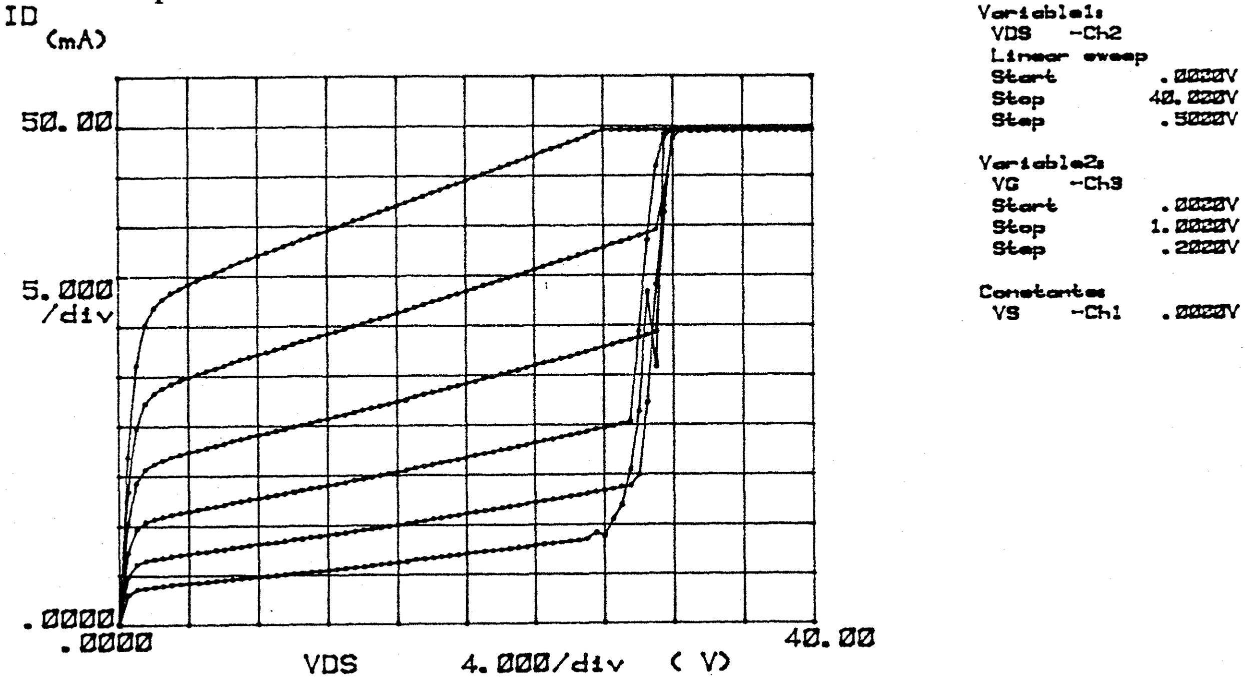
6.1.1. Vertical Power DMOS

Fig. 6-1 is a trace of the drain current verses the drain to source voltage with various applied gate voltages (G1). Fig. 6-2 shows the same device as Fig. 6-1 in the linear region of operation, at low drain voltages. The on-state resistance is determined directly from the slope of the line at $V_{G1} = 10V$. Fig. 6-3 is the characteristic trace of another device which did not break down or exhibit high leakage. This was the exception, most of the devices exhibited leakage currents and a lower breakdown voltage than expected (Fig. 6-1), as well as a high on-state resistance (Fig. 6-2).

The characteristics made it impractical to perform a transient analysis. Transistor action was certainly observed, although no single device had low enough leakage and a properly working driver. However, it was possible to measure the oxide capacitances. A complete C-V analysis could not be performed because of the leakage on the material, so the static capacitance measurements were compared to the calculated values.

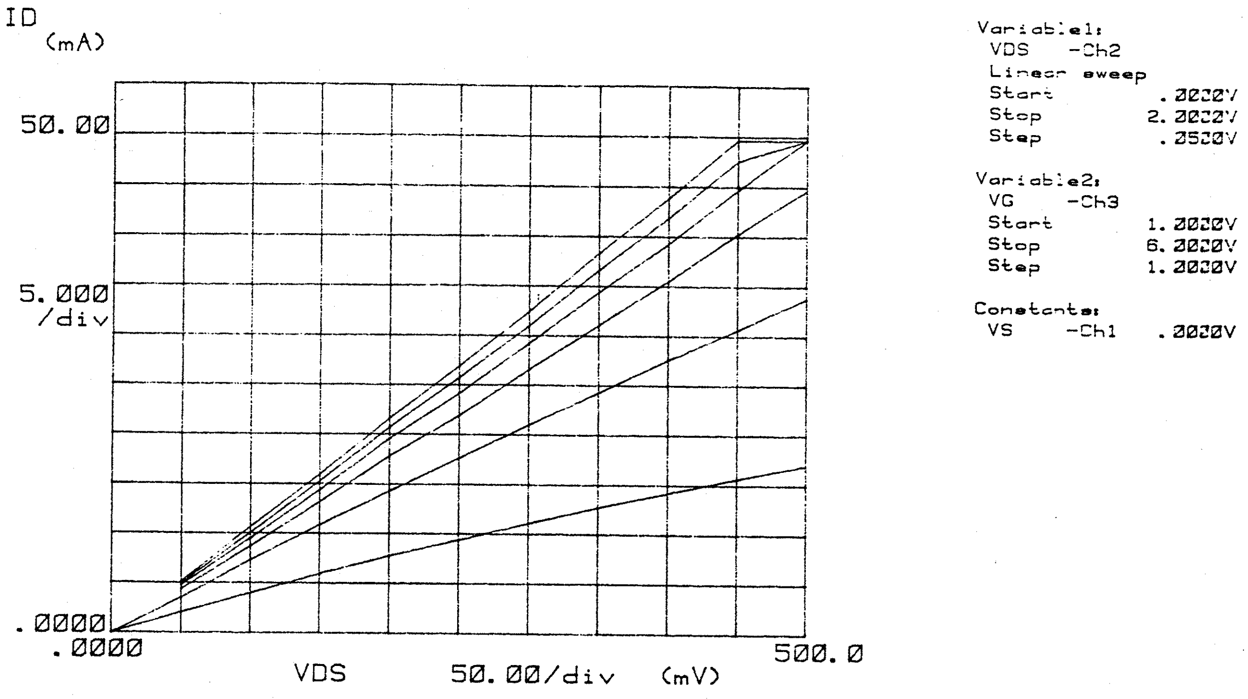
Capacitance	Calculated (pf)	Measured (pf)
G1 to S	212	240 ± 20
G2 to S	14	18 ± 3
G1 to D	240	190 ± 30
G1+ S to D	340	350 ± 60

The close agreement shows that the geometrical assumptions are generally valid. The tests were made at 1 MHz with no applied bias. The signals were noisy since a good contact could not be made to the chuck. Several devices were tested resulting in a spread of values.



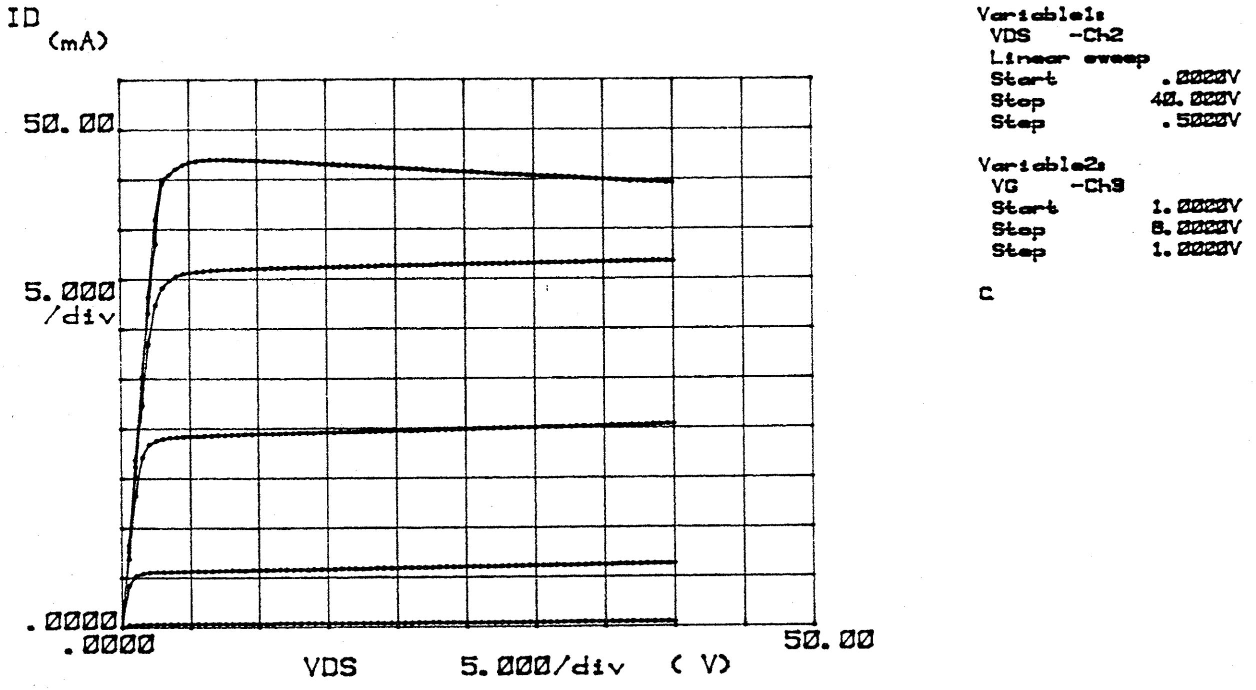
```
Variable1:
 VDS    -Ch2
 Linear sweep
 Start            .0000V
 Stop           40.000V
 Step             .5000V
Variable2:
 VG     -Ch3
 Start            .0000V
 Stop           1.0000V
 Step             .2000V
Constants
 VS     -Ch1      .0000V
```

Fig. 6-1 Transfer characteristics for the fabricated vertical DMOS measuring drain to source voltages up to 40 volts.



```
Variable1:
 VDS     -Ch2
 Linear  sweep
 Start            . ØØØØV
 Stop            2. ØØØØV
 Step             . Ø5ØØV
Variable2:
 VG      -Ch3
 Start           1. ØØØØV
 Stop            6. ØØØØV
 Step            1. ØØØØV
Constants:
 VS      -Ch1     . ØØØØV
```

Fig. 6-2 Same as Fig. 6-1 at low drain to source voltages.



```
Variable1:
 VDS     -Ch2
 Linear  sweep
 Start            . ØØØØV
 Stop            4Ø. ØØØV
 Step             . 5ØØØV
Variable2:
 VG      -Ch3
 Start           1. ØØØØV
 Stop            8. ØØØØV
 Step            1. ØØØØV
C
```

Fig. 6-3 Transfer characteristics of a vertical device which did not break down.

6.1.2. Lateral Driver

The lateral devices were tested and found to have breakdown voltages between 10 and 20 Volts. The on-state resistance, however, could not be accurately measured because of the geometry of the metalization. The aluminum resistivity was found to be approximately 0.1 Ω/□ while the G1 metal is approximately 50' long. Hence, since the ON device should have a resistance of less than 0.75 Ω, it is not possible to accurately determine the total on-state resistance.

From Fig. 6-4, one can see that the lateral devices exhibit transistor action and they also have high leakage. The on-state resistance, as determined from the slope of Fig. 6-5, is approximately 9 Ω, which means the true device resistance is more than 3Ω. This is not acceptable for use as the driver, so these devices are not of sufficient quality to warrant performing a transient analysis.

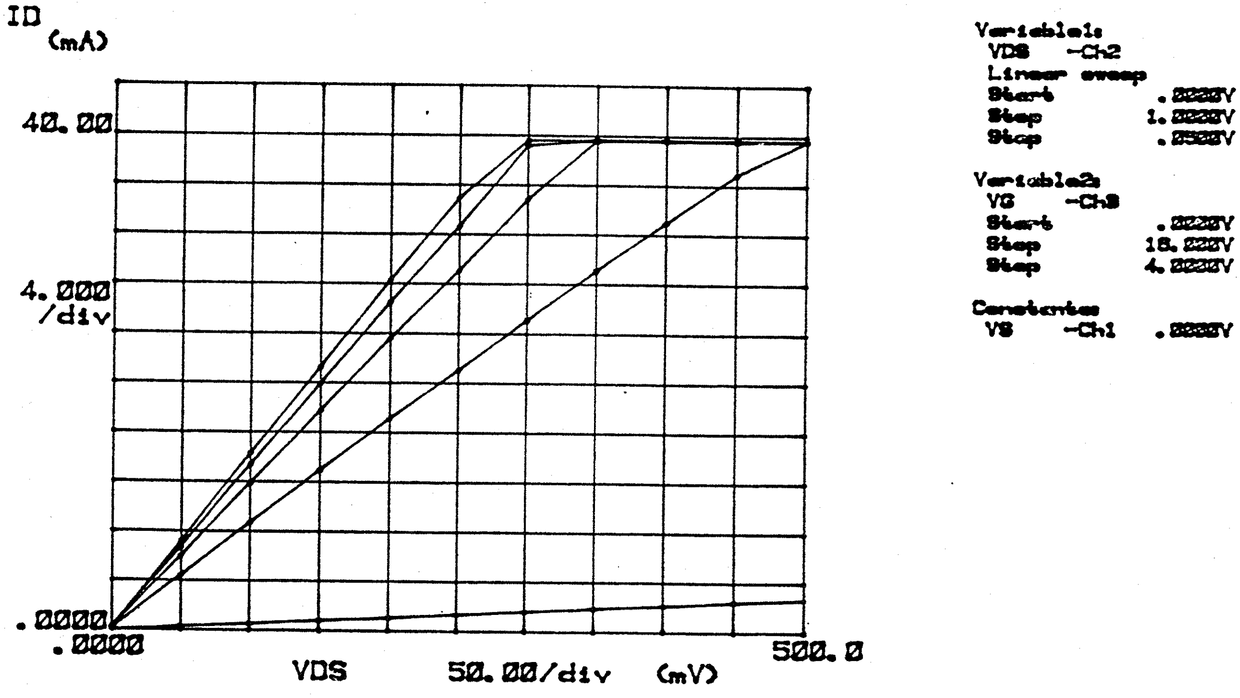
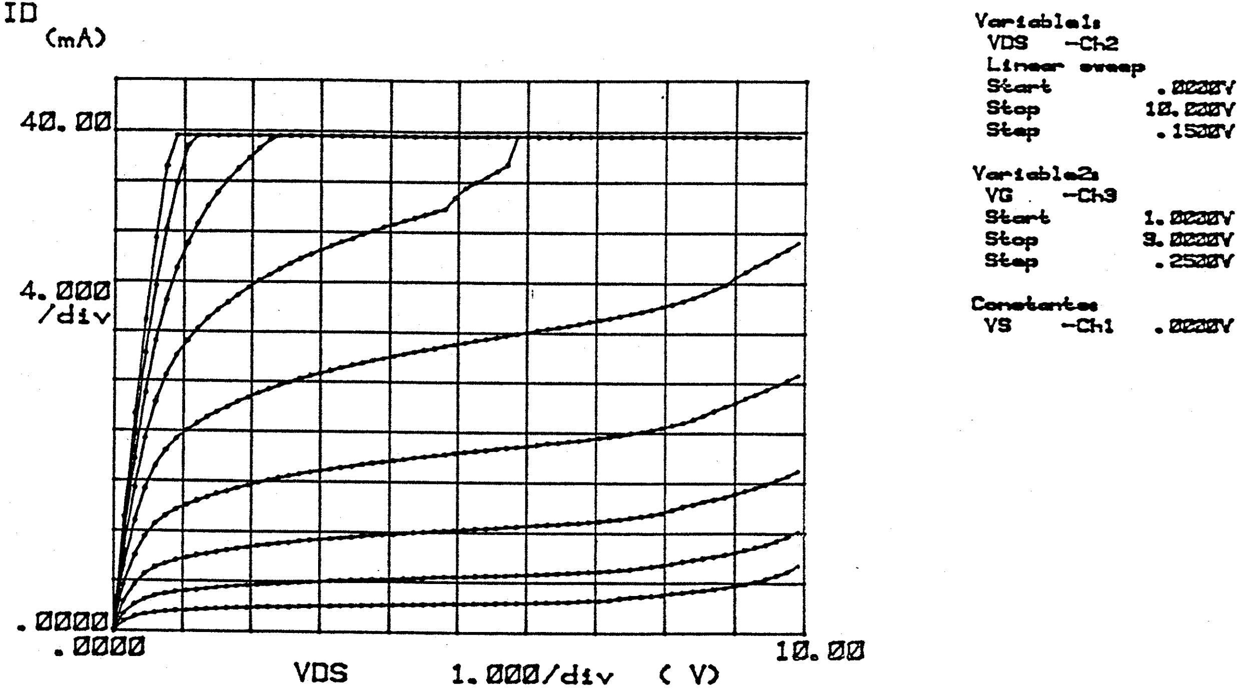


Fig. 6-4 Transfer characteristics of a lateral MOSFET driver



```
Variable1:
 VDS    -Ch2
 Linear sweep
 Start          .ØØØØV
 Stop          1Ø.ØØØV
 Step           .15ØØV

Variable2:
 VG     -Ch3
 Start         1.ØØØØV
 Stop          3.ØØØØV
 Step           .25ØØV

Constants:
 VS     -Ch1    .ØØØØV
```

Fig. 6-5 Same device as in Fig. 6-4, but at low drain to source voltages.

CHAPTER VII.

CONCLUSIONS AND RECOMMENDATIONS

From the characteristic traces, some conclusions can be made about the structure which will improve the fabrication of these devices in the future. Chapters 4 and 5 describe the finalized sequence based on observations during the fabrication process and on considerations for making the masks more reliable. Problems have been determined and rectified, thus, following the process sequence (Appendix) will produce reliable devices.

The coincidence of higher on-state resistance and lower breakdown voltage than expected for the epi thickness and doping suggests that there may be a large density of crystalline defects in the epi material. This hypothesis is further supported by the high leakage currents observed with zero gate voltage. The threshold voltage does not depend on the leakage, which suggests that the current results from the junction and not the channel. Crystalline defects in the junction act as generation sites, causing a high reverse current and a low voltage junction breakdown. These same defects, if distributed throughout, will decrease the electron mobility, and raise the resistance the epi.

Since the quality of the wafers had been questioned, some unused epi wafers were carefully inspected. Under close observation, the wafers were found to have ridges across the surface. These visible defects suggest that there are slip dislocations and line defects throughout the material, and they will certainly cause the types of problems observed. These wafers were then found to be over 11 years old, and their history is unknown. Consequently, all the irregularities in the transfer characteristics could be explained by the poor starting material.

Transistor behavior was observed on most of the devices tested, so one can be reasonably sure that the process itself is sufficient and would produce reliable devices if better starting material is used. Thus, for the next set of wafers to be produced, new high quality epi has been bought, and a new set of masks was designed on order to fabricate Power DMOS with integral shorting switches to be used in the power supplies being designed at MIT.

7.1. What's next

Having completed the structures and tested the devices, there are several improvements which can be made with respect to the design. These improvements include: using the self-aligned p+ contact opening to also self-align the metal, lifting the gate off the center of the neck region between p-wells to reduce the miller capacitance, and using a metal or silicide gate to reduce the series gate resistance. The first two will reduce the capacitance in order to allow for higher frequency operation, while the last suggestion may reduce the resistance so much that integrating the driver may not be need to be interdigitated, but may be placed in a seperate section of the device area.

Fig. 7-1 illustrates the way that two oxide etch masks can be used to align the metal to the source with one mask, while the second mask aligns the contacts to the gate and the shorting switch. Since two steps are needed for the final oxide etch anyway, the addition of a mask layer will not reduce the reliability. Also, the two masks align to different parts of the wafer, so their tolerances can be larger than the single oxide mask which must be stepped twice.

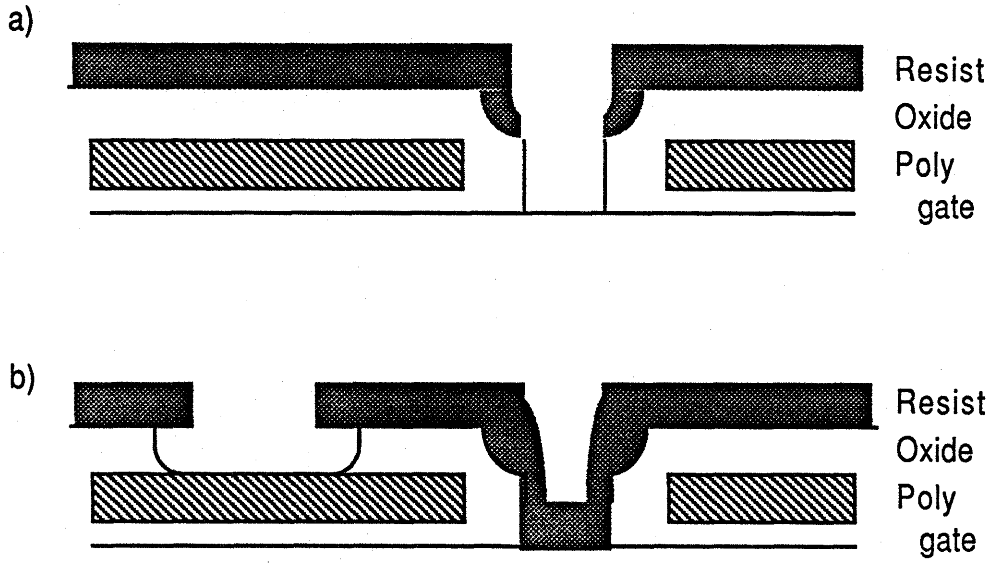


Fig. 7-1 Two oxide etch mask layers for self-aligning the metal to the source with the same openings for the p+ contact, (a) deifnes the openings to the source and (b) defines openings elsewhere and to the gate.

This modified process only requires a single oxide deposition step before the p+ contact is implanted. The wafer is then annealed using the first deposited oxide as the isolation layer. The oxide which must be etched by the first step is less than 1000 Å of thermal oxide, so it needs to be etched for less than 1 minute in BOE. The other layer (b), however, identifies openings in the 1.5 μm thick deposited isolation oxide and must therefore be etched for over 10 minutes. This modification can be easily implemented with nearly the same process sequence as the devices fabricated for this project. The only difference is that the first oxide deposition step must provide the entire 1.5 μm of isolation oxide, eliminating the second oxide deposition step.

In order to incorporate a lifted gate or a metal silicide gate, the process would have to be substantially modified. One possible way of achieving both of these is to grow a thermal isolation oxide before diffusing in the channel, define the source diffusions, and then depositing the silicide. Some refractory metal silicides can be processed at high temperatures so the same self-alignment techniques can be employed, however, if aluminum is used, one must be sure to develop a process which does not include any high temperature diffusion steps after the metal has been deposited. Since oxide can be deposited on top of metal, a two layer aluminum process is certainly reasonable, although, the gate will not be self-aligned to the sources.

Overall, the technology is very nearly at the point where 1 MHz switching circuits and 10 MHz resonant converters will be practical to produce on a large scale. Once the rest of the project has been completed, including improvements in the magnetic elements, power supplies will be much smaller and efficient than they are today. Consequently, in the very near future, power circuitry used in computers, or any large electrical system, will once again be only a small portion of the overall system's size.

REFERENCES

1. S.M. Sze, Physics of Semiconductor Devices. J. Wiley and Sons, New York, 1981.

2. S.M. Sze, editor, VLSI Technology. McGraw Hill Book Co. 1983.

3. R.A. Colclasser, Microelectronic Processing and device design. J. Wiley and Sons, 1980.

4. P.L. Hower and M.J. Geisler, "Comparison of Various Source-Gate Geometries for Power MOSFET's," IEEE Trans. on Elec. Dev. Vol. ED-28, No. 9, Sept. 1981 pg1098.

5. P.L. Hower, T.M.S. Heng, and C. Huang, "Optimum design of Power MOSFETs," IEEE IEDM Digest-1983, pg 87.

6. S.C. Sun, and J.D. Plummer, "Modeling of the On-Resistance of LDMOS, VDMOS and VMOS Power Transistors," IEEE Trans. on Elec. Dev. Vol. ED-27, No.2 Feb. 1980 pg 356.

7. M.H. Chi and C. Hu, "Some issues of Power MOSFETs," IEEE IEDM Digest-1982, pg 392.

8. R. Sittig and P. Roggwiller, editors, Semiconductor Devices for Power Conditioning, Plenum Press, New York, 1982.

Appendix. Detailed Process Sequence

This appendix describes the complete process sequence to yield usable devices. It assumes a certain familiarity with general silicon device fabrication, and should provide sufficient information for an experienced individual to produce reliable devices. Because of small photolithographic dimensions, the wafer stepper requires a great deal of patience to adjust the alignment, focus, and exposure. This means exposing at least one alignment monitor and one focus-exposure monitor before each device wafer run.

Start with (100) n on n+ silicon epitaxy wafers with good crystalline integrity, since poor starting material results in low yield and low quality final devices. Use 15 to 19 μm of 2Ω-cm n-type silicon on 0.002-0.005 Ω-cm Arsenic or Antimony doped substrates. Before beginning the process, visually inspect the wafers for gross contaminants or obvious defects. A typical run should consist of at least 4 device wafers and 6 monitor wafers of the same crystallographic orientation with higher resistivity than the epi. The monitors allow one to characterize each individual step during the process, so any problems should be detected immediately. All 10 wafers should be appropriately marked on the backside with a diamond scribe so they will not be confused with one another. The monitors will be described as M1 - M6 for reference.

A.1. Gate oxidation and poly deposition

These steps proceed immediately from one to the next. HCl clean the oxidation tube while RCA cleaning the wafers. Perform the HCl clean for at least one full hour at 1100 °C immediately before the oxidation. Also, be sure that the Polysilicon furnace will be calibrated for deposition soon after the oxidation. Work out the timing so that immediately following the RCA clean, the wafers go directly into the middle oxidation furnace, immediately followed by 1.1 μm of poly deposition. Once the cleaned wafers are spun dry, load all 10 wafers on the boat in the oxidation tube. Load them in a nitrogen environment at 1100 °C. After 5 minutes, turn on the Dry oxygen (the nitrogen should turn off) and let them sit for 46 minutes. Then, turn off the oxygen (nitrogen should turn back on) and let the wafers sit another 15 minutes before they are removed. Preset the furnace temperature back down to "idle" when finished.

If the Polysilicon LPCVD system is not ready for the run, bring the wafers to the mouth of the oxidation furnace where they will sit at a low temperature but still be in the clean nitrogen flow of the furnace. Once the poly furnace is ready, transfer all but one monitor wafer (M1) to the poly furnace. After the deposition, remove and prepare the device wafers along with 2 monitor wafers (M2 and M3) for lithography. Measure the poly thickness of the monitors (M4 -6) and save them for determining the etch rates of later steps.

Measure the oxide thickness of the first monitor wafer (M1) and record the value. Evaporate aluminum over the wafer through a shadow mask, making MOS capacitors. Evaporate metal on the backside as well, and anneal it in forming gas at 450 °C for 30 minutes. Perform C-V measurments to determine the oxide quality and surface state density. This wafer, and all completed monitor wafers, should be kept along with the run for future reference.

A.2. Photolithography #1

The preparation for photolithography for this layer repeats for every mask step. First, put the wafers in individual tins, and place them in the "Dehydration bake" oven for 30 minutes at 200 °C. Make certain that the syringes contain sufficient HMDS and Positive 1370SF resist for coating all 6 wafers. Check the speed and timer on the spinner for 5000 rpm and 30 seconds respectively. After the wafers come out of the oven, place them on the spinner. Coat each wafer with HMDS and spin it. Wait about 15 seconds and coat them with 1370SF positive resist and spin again. Remove the wafer from the spinner chuck and place it back in its tin. Let the wafer air dry until all the rest of the wafers are spun. Then wait an additional minute or so, and place all the wafers in the pre-bake oven at 90 °C for only 25 minutes. Be sure to remove the wafers after that time because extra heat reduces the photosensitivity of the resist.

The wafers are now ready for exposure. Be sure to schedule time on the wafer stepper within the next two or three days. During this time, wafers should be stored in a cool dry location. The GCA environmental chamber is an ideal location for storing the wafers before they are exposed.

Using the retical DP (mask#1), set up the wafer stepper for the mask parameters. Close the apertures around the retical to the appropriate size of the emulsion. Open a file for the run. All the masks use the same set of alignment markers, so the same file is used for all five masks. The CAD layout program gives the values for the key offsets and the stepping distances. These values are recorded in the file on the wafer stepper.

A focus/exposure run made with one of the monitor wafers determines the proper settings for the rest of the run. Values from previous users offer a good point around which to center the test. Expose the wafer and develop it in a 1:1 solution of 312 developer and DI water. Observe the resulting pattern under a microscope and determine which die has the clearest image and the sharpest lines. Once determined, type the row and column of that die, and the computer informs the operator of the selected focus and exposure setting.

Next, all 5 wafers may be stepped through without alignment. Keep track of the monitor wafer M3 since it serves to check the alignment of the rest of the layers. Develop the wafers and post-bake them for 30 minutes at 130 °C. The wafers should be individualy inspected for resist quality. If any of the wafers look questionable, the resist should be stripped and organically cleaned, then the photolithography sequence should be repeated for that wafer.

A.3. Polysilicon etch #1

Immediately before etching the poly, prepare a solution of buffered oxide etch (BOE). Etch the oxide on all 5 exposed wafers and one of the poly monitor wafers (M4). Etch the wafers until the solution sheaths off the silicon assuring that no more oxide is on top of the silicon. Set the etch for 4.5 minutes using the standard polysilicon recipe, and use the end point detector connected to an X-Y sweeping plotter. Place the poly monitor wafer in the system and etch. When the end point drops, notice the time remaining on the video screen. This time should be around 45 seconds, meaning that the etch took 3 minutes 45 seconds. Set the over-etch to 10% to make sure the poly is completely removed. Run all the device wafers with the same time as determined from the monitor, each with an over-etch. The endpoint may be used to observe the process, although the automatic detector will not necassarily observe the completion of the poly etch because of the low resolution with the pattern.

A.4. P-well implant and diffusion

Send all 4 device wafers and the etched poly monitor from step 3 (M4) for implantation of boron with a dose of $5x10^{13}$ cm^{-2} at 50KeV. Keep the etched monitor wafer (M3) aside for testing the alignment of following lithographic steps.

Do not strip the resist off the device wafers until after they are implanted. Upon the return of the wafers, use the oxygen (organic removal) plasma on the Day etcher. Etch the device wafers and the two monitor wafers which still have resist (M2 and M3). Apply the plasma until the color of the plasma changes from white to mauve. Then remove the wafers from the plasma and perform a complete RCA clean to the device wafers and M4 in preparation for the drive-in. While there is no need to perform an HCl furnace clean before the drive-in, the furnace should be heated to 1150 °C with dry O_2 flowing through it at least 15 minutes before the wafers are deposited.

Load the wafers on the boat of the lower oxidation furnace immediately following the spin drying. Place the boat in the furnace with oxygen still flowing through the tube. Move the boat to the center of the furnace and close the tube cover. Immediately discontinue the oxygen flow, making sure the nitrogen is running properly. These wafers need to diffuse for 6 hours, so the clean should be scheduled early enough in the morning so they can be removed at a reasonable time of day. Once completed, the wafers should be slowly removed from the furnace and allowed to cool. Reset the furnace temperature to its idle value.

A.5. Source implant

Send all 5 of these wafers for a second implantation. Prepare them for implantation of the heavy n+ dose. They need no further preparation once removed from the furnace. They require an arsenic dose of 10^{16} cm^{-2} with an energy of 150KeV. Once they return, prepare them for a second photolithography as in step 2.

A.6. Photolithography #2

This time there will be an alignment, so both the plain poly wafer (M2), and the alignment wafer (M3) need to be prepared, as well. Follow the procedure for spinning resist on all 6 wafers. Set up the mask PG for the poly removal around the devices. Perform the focus/exposure test to M2, then expose the alignment wafer. Under a microscope, determine that the alignment is correct. If not, adjust the key offsets appropriately. Expose the remaining 4 device wafers, develop, and post bake them. Then follow the same proceedure for plasma etching the poly as in step 3, using a second monitor wafer (M5).

A.7. Polysilicon etch #2

Following the post bake, sign up for using the LAM etcher. Afterwards, the resist must be removed from all the wafers. Etch them in the oxygen plasma with the Day etcher for approximately 5 minutes, or until the plasma changes color. Then RCA clean the wafers M2 and M5 along with the 4 device wafers in preparation for the deposited oxide.

A.8. Oxide deposition #1

Having cleaned 6 wafers, place them in the PYROX CVD oxide deposition system. The wafers must be cleaned just before placing them in the system, so arrange the time with the operator accordingly. Deposit 1µm of oxide at the standard temperature of 450 °C. Depositing slightly more than 1µm of oxide is not a problem in this step. Since this oxide only serves as a mask for the implantation and to align the plug removal, this oxide does not need densification. Therefore, follow the procedure of spinning resist on 5 of these wafers (all but M5) and on the alignment wafer (6 wafers in all) and prepare the third mask DD in like manner to step 2.

A.9. Photolithography #3

Using the same procedure as step 6 and 2, perform a focus/exposure run with M2, followed by an alignment test with M3. Again, be sure to adjust for alignment errors after exposure of this wafer. Use the mask DD for removal of the poly for p+ diffusion. Since this step etches the deposited oxide over poly, use the monitor from the previous step for focus/exposure. Strip the resist with acetone before proceeding so it can again be used to monitor the oxide removal.

A.10. Polysilicon etch #3

Strip the oxide from the 4 device wafers and the monitor from the previous step (M2). Etch them in BOE for approximately 5 minutes, until the oxide completely sheaths off the monitor wafer and etch the monitor M6 for about 10 seconds to remove the native oxide. Then, use the LAM etcher to remove the poly. Following step 3 for the final time, with the monitor M6, set the etching time, and etch the device wafers in like manner to the other poly etch steps. Keep these 5 wafers together for the final implantation.

A.11. P+ implant

Send all 5 wafers from the previous run for implanting a boron dose of $5x10^{15}$ cm^{-2} at an energy of 200KeV. Upon their return, strip the resist off all 4 device wafers and monitors M2,4, and 6. Etch them in 10:1 HF with M5 until all the oxide is removed from the monitor. Bring the other monitor used for the first and second implants together with these and RCA clean all 8 in preparation for the final oxide deposition and drive-in.

A.12. Oxide deposition #2

The final oxide deposition occurs in a sequence of steps which should be completed in the same time period, one after the next. It can all be performed in the same day. Following the oxide removal of the 6 wafers driven-in from the previous step, RCA clean the wafers and deposit 0.5 µm of oxide on them as soon as possible. Then perform a second RCA clean, and place the wafers in the lower oxidation furnace for densification at 900 °C in an oxygen environment for 30 minutes. If the PYROX system is ready for the second deposition, another RCA clean will not be needed. The wafers can go straight from the furnace to the CVD system. This time, deposit 1µm of oxide. Then RCA clean the wafers in preparation for the final long diffusion.

A.13. Diffusion #2

Diffuse all 6 wafers in the lower oxidation furnace at 1000 °C for 15 hours. Be sure to sign up for the wafer cleaning station late enough in the day so they may be retrieved at a reasonable time the following day. In similar fashion to step 4, run oxygen in the furnace at the process temperature before loading the wafers, and change over to nitrogen immediately after they have been loaded. The whole diffusion takes place in a nitrogen environment to avoid oxidation.

Sign up for two time slots on the PYROX CVD oxide system. Before using that station, the oxide must be removed and an RCA clean must be performed. Schedule an appropriate time to accommodate the following sequence of steps. Using a shadow mask, make MOS capacitors on the monitor M5 to determine the quality of the isolation oxide.

A.14. Photolithography #4 and #5

The double oxide etch procedure avoids excessive undercutting. First, follow the procedure of step 2 for spinning on the resist and soft-baking. Use the alignment wafers (M3) and the monitor M2 for focus/exposure. Set up the oxide mask DX for defining the contact vias. Once again, perform a focus/exposure run, followed by an alignment run and strip the resist with acetone from the monitor wafers. Postbake the 4 device wafers, and etch them in BOE along with the two oxide monitor wafers M4 and M6 with no resist on them. Etch for only 8 minutes and remove these 6 wafers.

Repeat the entire sequence once again, being sure the alignment and exposures are correct. This time etch all 6 wafers until sheathing occurs with the monitors. They should only need another 7 or 8 minutes. Once all the oxide has been removed from the monitor wafers, inspect the device wafers under the microscope, making sure the oxide has been satisfactorily etched from the contact holes. The holes should look shiny and silvery, like silicon; not dull, or colored. If Oxide still is apparent, etch them more. Agitation may be necessary, but be careful not to over-etch, otherwise the metal may short to the poly.

A.15. Metal deposition

The wafers should be RCA cleaned with a 10 second dip in 10:1 HF immediately before the metal is deposited. After this clean, load the wafers on the e-beam evaporator chuck and have the operator deposit 0.75µm of aluminum. Use wafer M2 as a monitor again for the focus/exposure measurement. Then prepare all 5 aluminized wafers for the final photolithography step along with M3 for alignment.

A.16. Photolithography #6

Follow the procedure of spinning resist on the 5 wafers from the previous step and on the alignment wafer (6 wafers). Prepare the mask that defines the metalization, DM, on the GCA wafer stepper. Run the focus/exposure test followed by the alignment test and expose the 4 device wafers. Hard bake the 4 device wafers and prepare the PAN etch solution.

Etch the wafers, one at a time, in the PAN etch. Be sure the area is well ventilated and wear gloves to prevent exposure to the chemical. Etch the wafers for approximately 10 minutes, until the aluminum is completely stripped around the devices. Rinse each wafer well in DI water and inspect them under the microscope to assure that all the aluminum has been removed where it should be. If not, etch for as long as is needed until oxide appears between the metalization. Continue this procedure for all 4 wafers.

A.17. Backside preparation

Without removing the resist from the front of the wafer, spin a second layer of resist without HMDS and without a dehydration bake. Hard bake the wafers for 30 minutes. Then, place them in the silicon etch canister of the Day plasma etcher with the device side down. Run the SF_6 plasma on the backside at 300 Watts with a pressure of 100 mTorr. Observe the backside, and etch for about 7 minutes until the silicon is removed. The plasma will stop etching once the silicon is reached, so turn off the plasma at that time. Purge the lines with nitrogen, and remove the wafers.

Etch the wafers in BOE for under 2 minutes until the oxide is completely removed from the backside. Then, using the thermal aluminum evaporator, deposit aluminum on the backside of these wafers and etch off the resist from the front side using the oxygen plasma. Now the metal is ready for alloying. Place the wafers in the low temperature anneal furnace with forming gas (20% H_2, 80% N_2) at 450 °C for 45 minutes.

A.18. Testing the wafers

The process sequence is now complete. To determine the properties of the devices, measure the sheet conductivity of the two monitor wafers, and grind down a section to determine the junction depths. Record this information and compare it to what the process should yield. Finally, test the devices on the wafer, and observe all the desired characteristics.