Preface

In Igor Stravinsky's ballet "The Firebird", Prince Ivan captures the exotic Firebird in the magic garden of the sorcerer Kashchei. The prince releases Firebird when she gives him a feather which can be used to summon her in times of danger. The Prince falls in love with a princess held by the evil Kashchei, and eventually defeats the sorcerer and wins the princess' freedom with the Firebird's help.

It is fitting that much of this text was written to the music of this ballet. At once elegant, graceful, coordinated; and disjointed, unexpected, and misunderstood, "The Firebird" all too closely resembles the nuances of Silicon-on-Insulator technology.

As we complete this textbook, it is finally becoming clear that our industry no longer regards silicon-on-insulator technology as being in the same category as gravity-powered shoe air-conditioners and jet-powered surfboards[1]. SOI has been incubating for perhaps 20 years. Although the pressures of scaling have increased the urgency of its development, there are those who argue that it's behavior remains too unpredictable to be accurately modeled. Suffice it to say that SOI's potential performance rewards for enduring this complexity are substantial. There is no question that SOI is not an easy technology to learn and design with; using the technology requires a thorough, fundamental comprehension, and demands considerably more designer mental energy than its bulk CMOS predecessor did. The reader must remember, however, that

1. See "The Gallery of Obscure Patents" at *http://www.patents.ibm.com/gallery* on the World Wide Web.

although SOI might be difficult, conventional CMOS was also originally referred to as the S.O.B. (Silicon-On-Bulk, of course!) technology when first introduced.

Chapter 1 provides the reader with a brief review of how CMOS technology has been scaled through multiple lithography generations, and describes the device physics which threaten conventional CMOS' future. The case for SOI is built as a means to continue improvement in density, performance, power and integration.

Chapter 2 introduces the basic SOI physical device structure, and how it differs from its bulk CMOS predecessor. Manufacturing considerations and materials properties are highlighted in this section.

Chapter 3 addresses the fundamental electrical properties associated with the SOI device, and how the structures described in the previous chapter give rise to these properties. If the reader wishes to "cut to the chase," this chapter and the following two provide the most essential of SOI core concepts.

Chapter 4 explores the response of static circuits to SOI technology and offers guidance on static design practice.

Similarly, Chapter 5 provides an introduction to the issues associated with dynamic circuits in SOI, and includes design practices disclosed in the recent literature which mitigate many of the undesirable dynamic problems.

SRAM and cache arrays are probably the hardest designs to move into SOI. Chapter 6 is entirely devoted to addressing the issues surrounding SOI SRAMs. A few suggestions are offered on how to get around some of these problems.

In addition to the SRAM, selected other special function circuits require some finesse when migrated into SOI, and are described in Chapter 7. Off-chip-drivers, phase-locked-loops, and Electro-static Discharge devices are among these unique structures.

Chapter 8 addresses global design considerations which are unique to SOI. Power distribution, clock branching, and noise is discussed.

Finally, Chapter 9 offers a glimpse into some of the future leverage SOI may offer. Rather than designing to avoid SOI's idiosyncrasies, the true benefit of the technology may be realized when these features are exploited.

CHAPTER 2

SOI Device Structures

2.1 Introduction

Silicon On Insulator (SOI) structures do not vary much from normal bulk CMOS. The major difference is the insertion of the insulation layer beneath the devices. Once this is accomplished, one could continue to use the identical bulk CMOS process and fabricate the devices. No changes to the process would be required, however, to take full advantage of SOI; small changes to the process are required. This chapter will discuss the physical structures and fabrication techniques of the wafer, FETs, diodes, resistors and thin oxide capacitors. Intertwined with the physical descriptions will be the mention of process changes to enhance the SOI device's usability and performance.

Another process for isolating FETs from each other is silicon on sapphire (SOS). This technique used a sapphire substrate and etched small mesas of silicon on its surface. The FETs on SOS can be made similar to the FETs on SOI. This technique was expensive to use since it required sapphire substrates and was not very manufacturable. The methods described in this chapter will use techniques that are compatible with today's manufacturing process for bulk CMOS wafers.

The subject of this chapter is the basic structure of devices found in SOI CMOS technologies. This includes the wafer structure, FET cross-sections, diode profiles, decoupling capacitor structures and resistor topologies.

2.2 *Wafer Fabrication*

So how do I get started?

The objective is to create a insulating layer beneath the devices such that the body of the FETs are not connected when the shallow trench isolation (STI) is used to pattern the active silicon areas. This insulating region is typically made of silicon dioxide and is formed one of three techniques: SIMOX, Bonded SOI or Smart Cut. This insulating region will be called the buried oxide or BOX.

2.2.1 SIMOX

SIMOX stands for the Separation by the **IM**plantation of **OX**ygen. In this technique, an oxygen implant is performed on an epitaxial wafer before it has started any other CMOS processing.

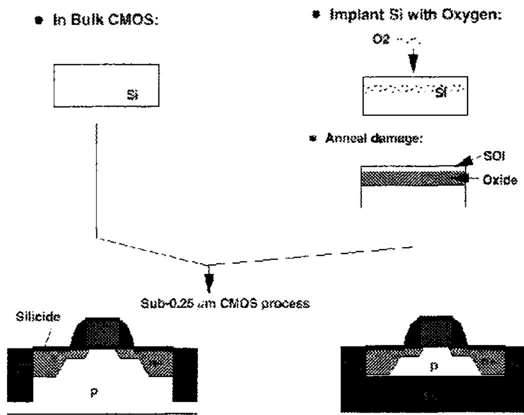

FIGURE 2.1 SIMOX process steps

The high energy implant forces a large dose of oxygen deep beneath the surface. Once the implant is complete, a high temperature anneal is done to form the silicon dioxide insulating region. The anneal must also help to recrystallize the silicon in the epitaxial layer which is at the surface of the wafer. Due to the high energy and high dosage of the implant, the surface of the wafer has been heavily damaged and silicon atoms

have been pushed deeper into the wafer causing a surplus of Si atoms beneath the surface. The post-implant anneal must recrystallize the surface to produce high quality silicon for the subsequent CMOS processing. During the recrystallization anneal, the volume expansion of the oxygen converting to SiO_2 results in additional mechanical stress to accumulate at the Si/SiO_2 interface. This may result in forming silicon dislocations or pipes that are 0.2 to 1.0μm long, see Figure 2.2. These dislocations are long

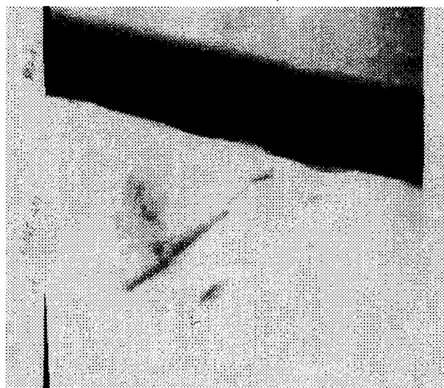

FIGURE 2.2 Dislocation defect in the silicon layer after recrystallization anneal.

enough to create a leakage path that may from source to the drain of an FET if one lines up with the length of the FET.

The dose of the implant required is dependent upon the device design point that is required for the given product. Too large of a dose is slow and expensive and will cause more defects in the top layer of silicon. Too small of a dose will not reduce the junction capacitance enough and will not electrically remove the silicon beneath the buried oxide from the active device. A typical dose for SIMOX is 10^{16} to $5e10^{17}$ cm^{-3}[2.7]. Therein, the buried oxide may be considered another FET oxide with the backside silicon as the gate. This will be discussed more in Chapter 8.

The buried oxide varies in thickness from 5nm [2.1] to 400nm [2.5] with varying doping. The energy of the implant will determine the silicon thickness above the buried oxide. Thicknesses of 50 nm [2.2] to 180 nm [2.6] are used for typical microprocessor designs. Other applications, such as space applications or power devices, will use much thicker active silicon regions. The thickness of the silicon layer is more related to the design of the FET. If one chooses a partially depleted FET instead of a fully depleted FET, the silicon layer is going to be thicker.

As with any added process step, the creation of the BOX is not for free. It requires a high energy implanter than can supply a very large dose of oxygen. Currently, about 20-40 wafers per day are created with this technique on a single implanter. As lower doses are applied, this number will increase, but this step is still a time consuming process. Another artifact of the implantation is that not all of the silicon is converted to SiO_2. When this occurs, small locations of silicon reside in the BOX. Due to the high dose, the occurrence of silicon residuals is most likely to occur at the bottom of the BOX and not near the device. These residuals are called inclusions.

2.2.2 Bonded SOI Wafers

Bonded SOI wafers create the buried oxide without ion implantation. Figure 2.3

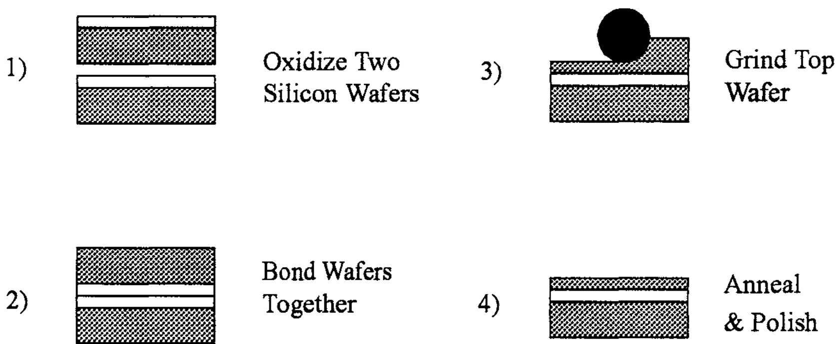

FIGURE 2.3 Process steps for SOI wafers created using the bonded wafer technique.

shows the processing steps for bonded wafers. First, two silicon wafers are heated to form a silicon dioxide layer on the surface. The two oxide surfaces are then bonded together to form the buried oxide. The backside of one of the silicon wafers is ground down to the desired thickness. Finally the SOI wafer is annealed and polished to leave a thin layer of silicon above the buried oxide that is electronic grade quality[2.7]. This wafer is then ready for typical CMOS processing. One problem with bonded wafers is that it required the use of two silicon wafers to provide one SOI wafer.

2.2.3 Smart Cut

The final technique discussed here is called Smart Cut. Figure 2.4 shows the steps in

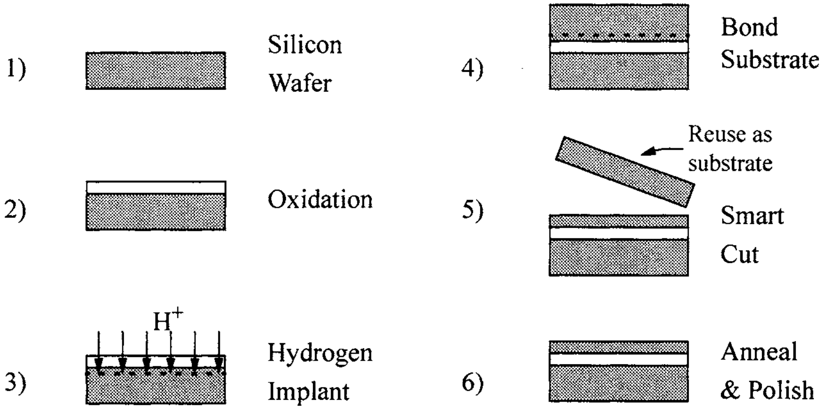

FIGURE 2.4 Process steps for Smart Cut.

the smart cut process. This technique starts with the oxidation of a silicon wafer. The wafer is implanted with protons (hydrogen nucleus) through the oxide layer. The surface of the oxide is clean. A second silicon wafer is bonded to the oxide layer (step 4). This wafer is now the substrate for the SOI wafer. A thermal anneal creates a stress fracture along the plane of the hydrogen implant. This is the Smart Cut. The original silicon wafer can be removed from the trilayer stack leaving behind a thin layer of silicon on the top of the buried oxide. The removed portion of the silicon wafer will become the substrate for another smart cut wafer. In this manner, no silicon is wasted. To finish process, the SOI wafer is annealed and polished to prepare the surface for traditional CMOS processing steps.

2.3 Patterning SOI regions

What if I only want some SOI circuits?

It is possible to pattern the region that is dedicated to bulk CMOS separate from the region that will be SOI. The patterning techniques are most easily applied to the

SIMOX process since one would be patterning the implantation region. Both bonded SOI and smart cut would require an alignment of the patterned areas to each other. The transition region between native bulk silicon and SOI regions is dependent upon the thickness of the buried oxide. This transition region is not usable for device processing due to the stress that exists in this region. Figure 2.5 shows an example of a transition region. The addition of the oxygen from the implant and the subsequent anneal creates a small thickness difference between the bulk and SOI region. This thickness difference is highly stressed and can result in cracking of the silicon at the surface (See Figure 2.5). Transition regions are in the tens of micrometers range for a

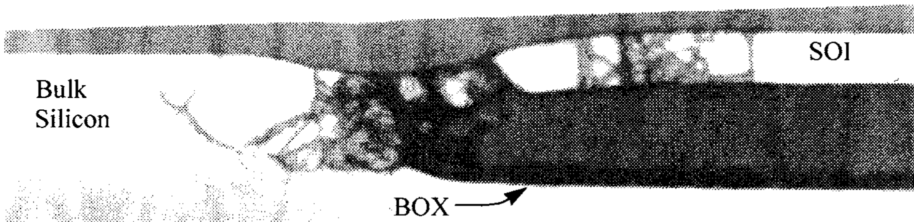

FIGURE 2.5 TEM of the transition region from bulk to SOI (Complements of IBM Research).

2000 μm thick buried oxides. As the buried oxide thickness become smaller, the transition region can become less than 5 μm. Therefore, if one desires to have bulk circuits on one part of the chip die and SOI circuits on another part of the die, it is feasible, though not regularly practiced.

One should not take too lightly the small step that occurs in the transition region. This height difference can be a real challenge for photolithography, trying to pattern the next several levels. The focus depth of today's lenses are very shallow and cannot focus at two different depths for the bulk and SOI portions of the wafer.

2.4 Transistor Structures

I have a buried oxide, now what?

The MOSFET is the structure under scrutiny and research in both bulk CMOS and SOI CMOS. In bulk, there is only one main device structure and the challenge presented to device engineers is how to make the fastest, most reliable device. The

choice of thresholds voltages, junction depths and background doping levels all have driven scaling of the FET to the to ever-finer device dimensions.

With SOI, a device engineer needs to make one more basic decision before selecting the rest of the device profiles and parameter. The question to be answered is: Are you going to work with a partially depleted device or a fully depleted device. As with most engineering questions, this one dictates or is dictated by other manufacturability questions such as control of the threshold voltage, the silicon thickness, etc. Let's look at the device structures.

2.4.1 Partially Depleted FET Structure

Keep your body in control!

Relying upon one's understanding of a bulk CMOS FET, the SOI FET has many features in common. Figure 2.6 shows the cross-section of the partially depleted SOI

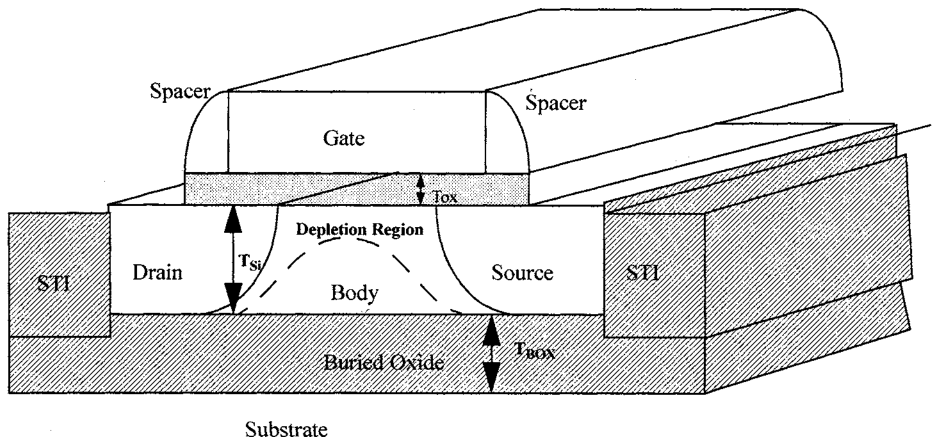

FIGURE 2.6 Cross-section of a partially depleted SOI FET.

device. This figure is not drawn to scale. The silicon is insulated with silicon dioxide on all sides. Beneath the silicon is the buried oxide that was created before any device processing was started. To the left and the right of the active silicon, the shallow trench isolation (STI) is the same STI isolation technique that is used in bulk CMOS technologies. The shallow trench isolation is greater than the thickness of the silicon, T_{Si}, resulting in a continuous silicon dioxide region from the wafer surface through

the buried oxide beside the FET. The gate oxide thickness, T_{ox}, is identical to a bulk technology as is the gate stack thickness and the spacer dimensions.[1]

The depletion region extends into the body of the FET under the gate at the source-body and drain-body junctions. It does not deplete all of the charge in the body, resulting in the name *Partially Depleted* SOI.

One cannot control the voltage of the body with out having some region of the body containing charge. With the partially depleted body, there is a high resistance across the body, but the charge is mobile. The series resistance down the length of the body is large and variable, as the depletion width is voltage dependent. As the depletion width from the source or drain into the body increases with voltage bias, the cross sectional area containing charge becomes smaller and the resistance increases.

This mobile charge results in unique SOI-only device and circuit characteristics that will be described at length in several subsequent chapters.

This structure has a parasitic bipolar device in parallel with the FET. For an NFET, the parasitic bipolar device is an NPN structure. Without a contact to the body, the body is floating and under DC conditions is driven to a potential determined by the leakage currents from the source and drain to the body across the diodes that form the junctions. Under AC conditions, the diodes between the body and the source or drain, the impact ionization currents from the FET and the capacitively coupled currents from the other three terminals determining the potential on the body. When the body is floating for the FET this also means that the base of the parasitic bipolar device is floating. One wants to minimize the potential in the base region to prevent the bipolar device from turning on. In a similar structure, there is a parasitic PNP bipolar transistor that is in parallel the a PFET. The bipolar devices do not have very high gain since the base of the NPN or the PNP is quite wide. When the body is floating, there is no direct connection to the base of the bipolar transistor to sustain the potential.

---

1. A fully depleted FET structure is very similar to the partially depleted structure. It has the same STI isolation and it has buried oxide beneath the silicon. The one big difference is that the silicon thickness of a fully depleted device is thinner than the silicon thickness of the partially depleted device. The background doping of the NWELL is lower than in partially depleted devices. This allows the depletion regions from the source to body and from the drain to body regions to fully deplete the body of mobile charge under all bias conditions. The silicon thickness of a fully depleted device is very critical. This thickness determines the threshold voltage of the device and is a manufacturability issue. This text deals with partially depleted devices. See [2.3] for more information on fully depleted devices.

To lower the potential on the body, the source and drain region implants are done with a higher dose or a lower depth to increase the leakage across the diodes. With a more abrupt junction, the leakage helps to control the potential on the body.

As one looks from drain to source of the NFET in Figure 2.6, the dopant type is n-p-n. This structure creates a parasitic bipolar transistor. The drain is the collector, the body is the base and the source is the emitter. The base of the bipolar device is not connected to any potential. It is floating just like the body is floating. The parasitic device must be modeled and accounted for in device design points.

Another parasitic device that is apparent in Figure 2.6 is another NFET with the substrate acting as the gate and the BOX acting as the insulating oxide for the MOSFET. This device will be called the backside device.

Figure 2.7 shows an SEM cross section of a SIMOX wafer with tungsten local inter-

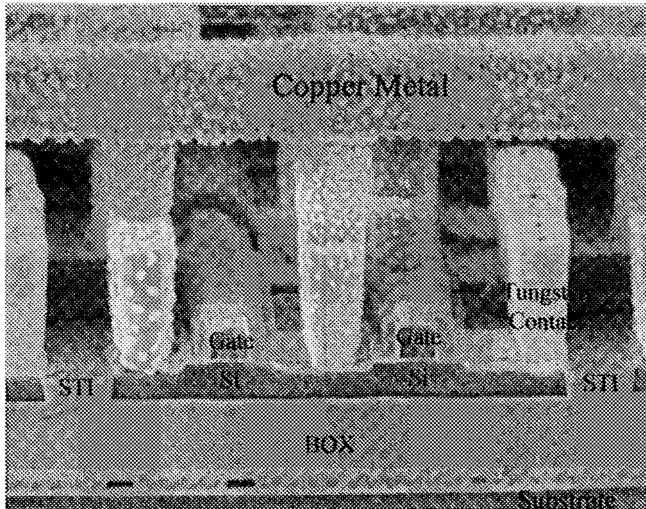

FIGURE 2.7 SEM of a cross section of two SOI FETs with tungsten contacts and copper wires.

connects, tungsten studs and one level of copper interconnects. The buried oxide (BOX) is much thicker than the SOI for this wafer. Here two FETs share a common silicon mesa and share the source connection between the two devices. The shallow trench isolation (STI) is present to the left and right of the devices and provides isolation from the other FETs nearby. The silicon above the buried oxide is surrounded by

silicon dioxide on all sides. The only places where the silicon dioxide is not present is at the contacts to the source and drain of the device.

2.4.2 Body Contacts

How do I control the body?

In bulk CMOS, it is possible to have independent control of the body of each and every device. However, it requires the use of a "triple well" process and much more space than SOI. In a more traditional twin well CMOS process on an p-type substrate, the body's of the PFETs are controllable by controlling the potential of each respective NWELL. There is no control of the body of an individual NFET since they share the common substrate which is usually grounded.

In a partially depleted SOI CMOS FET structure, there is a conducting path beneath the FET that can be used to control the potential of the body. This requires a new structure in SOI since each silicon area is an isolated mesa from the rest of the substrate or NWELL. To provide the conducting path from the body itself to a contact to the body, one must create a p-type channel for an NFET and an n-type channel for the PFET.

In Figure 2.8 two types of body contact are shown for NFETs. In part (a) of the figure,

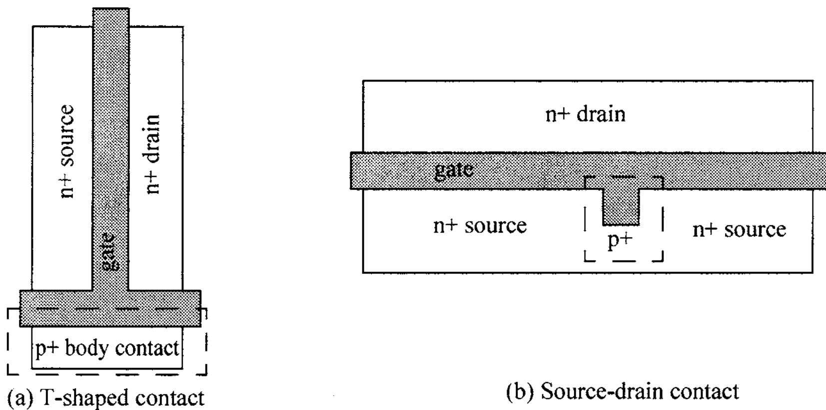

FIGURE 2.8 Body contacts for a partially depleted NFET.

the gate is a T-shape. A p+ region is created at the crossbar of the T-shaped gate. This p+ region is connected to the body of the NFET through the p-type region directly under the gate. Here, the body is a unique terminal and needs to be contacted with an

interconnect and a via at the body contact to control the potential of the body. One consequence of the T-shaped connection is that the width of the NFET is now increased by a long channel length device that connects the source and drain under the crossbar portion of the gate. Also, the resistance down the body can be quite large, and requires that the device be contacted at both ends. In this case the polysilicon gate becomes an I-shape. One may also finger wide devices by forming a multiplicity of smaller parallel gates and using one body contact to control all of the fingers. Finally, two separate devices may share a common body by using two T-shaped gates and aligning the crossbars opposite the common body. In this case, the body contact can be connected to an external node, or the devices may share body potential, but the body is not driven to an external voltage. This is useful in a circuit that needs to have matched devices such as sense amps in SRAMs.

In part (b) of this figure, an alternate contact to the body is achieved. In this method, the p+ region is created in the same silicon as the n+ source, creating a butted junction. Once a salicide is created on the source, the salicide will short the p+ region to the n+ source. The p+ region will again connect to the body through the p-type region that is under the polysilicon gate. There is no need to connect this p+ region to an interconnect since it is already connected to the source of the NFET. One disadvantage of this contact is that one cannot choose another potential other than the source to control the body. In some applications, it is desirable to have independent control of the body of the FETs. In this case, the T-shaped body contact is the structure of choice.

Additional body contacts can be derived. For example, the T-shaped structure in Figure 2.8(a) can become a body-source short if the left side of the crossbar is removed. Then the T-shaped gate becomes an L-shaped gate and the source and body are shorted again by the salicide. This will alleviate some of the space requirements required with the T-shape structure. In any manner, body contacts are not widely used as they increase the amount of capacitance on the gate, increase the charge within the body and take up more silicon area. Only critical circuits will have body contacts on them.

PFETs have symmetrical structures for their body contacts, but the contact regions are now n+ and connect to the n-type silicon. This n-doped silicon region is all that remains of NWELL beneath the polysilicon gate.

2.5 Diodes

Oh where, Oh where did my other terminal go?

With transistors, the fourth terminal of the FET comes into play with SOI much more than bulk. However, with simple p/n diodes, the substrate and the NWELL have been cut off due to the buried oxide. The removal of the NWELL has lost the second terminal on a diode in a bulk diode as shown Figure 2.9(a). One may consider using a butted junction, but this would require an additional step to prevent the formation of a salicide over the butted junction. A different topology of the diode can be created with the same processing steps currently in use.

2.5.1 Poly Bounded or Gated Diodes

In SOI, the diode can be constructed again, a cross-section is shown in part (b) of

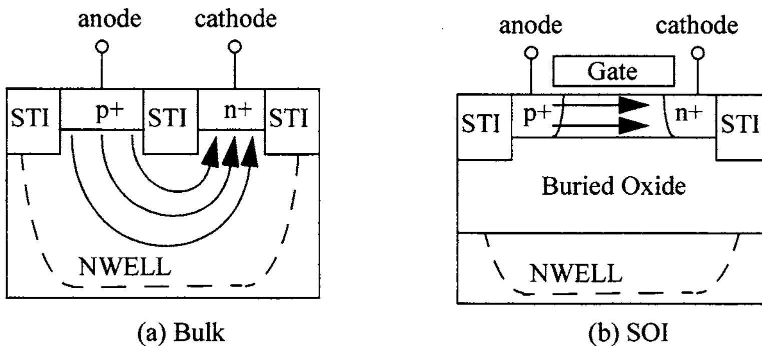

FIGURE 2.9 Cross-section of a diode in bulk(a) and a poly bounded diode in SOI(b) with direction of current flow shown by the arrows.

Figure 2.9. The cross-section looks something like an FET; however the diffusions that would normally be the source and the drain of the FET are now the anode and the cathode of the diode. The anode is p+ and the cathode is n+. The body of this structure could either be an NWELL (as shown), or the p- epi-layer. In this cross-section, the p/n junction is at the anode/body junction. The diode is now a perimeter diode, where the bulk diode was an area diode, meaning that the current flow was though the center of the diode. In the poly bounded diode, the current is only flowing through the edge of the diode. To pass an identical current through a forward biased diode, the

space required to layout a poly-bounded in SOI requires a larger space than a conventional diode in bulk CMOS.

The voltage on the gate determines the current carrying capability of the diode. One technique for using a gated diode is to force the gate to be connected to the cathode. This technique allows for the gate voltage to be maintained such that it does not effect the diodes current carrying capabilities. One should note that if the diode is used for an electro-static discharge protect device, (ESD Diode), on a multivoltage chip, the voltage of the output driver may cause the gate of the diode to exceed the reliability tolerance of the thin oxide. In this case, the gate should be driven to an intermediate voltage level that maintains the reliability of the oxide.

One complication of the poly bounded diode on SOI is that the series resistance is larger. The cross sectional area of the poly-bounded diode's NWELL is much thinner than the bulk diode's NWELL. This leads to series-limited current carrying capabilities of the diode. To correct for this, a much larger diode is used to reduce the total series resistance.

2.6 Resistors

Where is the path of least resistance?

The buried resistor and the polysilicon resistor are available in bulk as they are in SOI. The buried resistor is a depletion MOS device with its gate tied to the source. The resistance is determined by the channel region of the device, which remains unchanged. This results in the same resistance as well as the same tolerance in bulk and SOI technologies.

The polysilicon resistors are unchanged in SOI. As in bulk, they may be converted to silicide or left as polysilicon.

The NWELL resistor that is used in the bulk technologies is not operational in SOI since the buried oxide removes the silicon connection between two adjacent NWELL contacts. There is no conducting path through the NWELL in SOI. One of the other resistor structures must be used to replace NWELL resistors.

2.7 Decoupling Capacitors

I need charge.

The decoupling capacitor structure is nearly identical on SOI as it is for bulk CMOS. An accumulation capacitor, an NFET in an NWELL, is a simple structure for providing capacitance using the thin oxide of the FET. In this configuration, the source and drain are grounded, and hence the NWELL is grounded and the gate is tied to V_{DD}. As with the poly-bounded diode, the resistance of the body is large for the SOI decoupling capacitor and this results in the large RC component and reduces the capacitors frequency response. To alleviate this series resistance, the distance between the source and drain of the NFET should be kept short. This is a less-efficient use of space since the "length" of the FET is shorter than the bulk counterpart. Figure 2.10 shows the bulk and SOI decoupling capacitor structures.

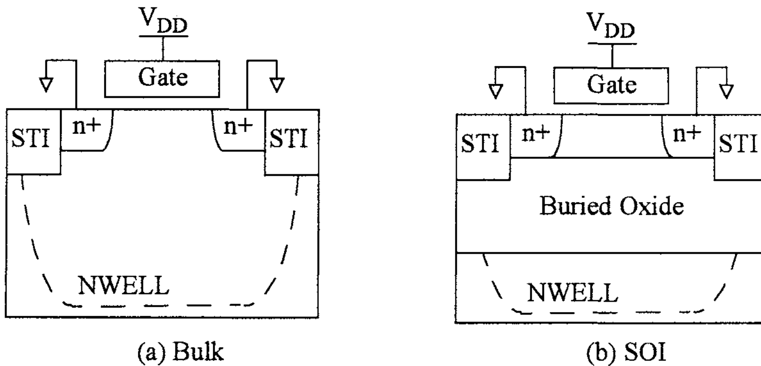

FIGURE 2.10 Cross-section of a decoupling capacitor in bulk (a) and SOI (b).

One problem with both of these structures is that a defect in the thin oxide will cause a V_{DD} to GND short and draw current. Common practice places a control FET in series between the source/drain node and GND. Therefore, if a defect does occur in the thin oxide, the control FET can be turned off and the current from V_{DD} to GND is eliminated. However, turning off the device removes the decoupling capacitance as well.

2.8 Summary

To make use of SOI, one needs to have available all standard active and passive device structures. This begins as soon as the SOI wafer is created. Three common techniques are used; SIMOX, bonded and Smart Cut. All three are capable of producing thin silicon regions on top of a buried oxide.

Once the wafer has been produced, the devices are needed. The FET device can come in two flavors: partially depleted and fully depleted devices. The main difference is how charge is kept within the body of the FET. Partially depleted devices do not have as stringent a tolerance requirement in the thickness as the fully depleted device. The body of the FET is floating in SOI. If one wants to control the body, body contacts are available at the expense of increased area and increased capacitance.

Other supporting structures such as diodes and resistors are feasible in SOI, but the basic structure is changed. Diodes require a poly gate to form the p-n junction and will have a higher series resistance. Buried resistors and OP resistors are unchanged, but NWELL resistors do not exist. Decoupling capacitors maintain the same layout, but will have more series resistance to the source and drain. SOI has removed the inherent decoupling capacitance that was present at the junction between the NWELL and the substrate.

With device structures in place, let's now consider the electrical properties of these devices.

REFERENCES

[2.1] ADVANTOX Spec sheet.

[2.2] P. F. Lu, et al., "Floating Body Effects in Partially-Depleted SOI CMOS Circuits", *1996 International Symposium on Low Power Electronics and Design, Digest of Technical Papers*, 1996, p.139.

[2.3] Neal Kistler and Jason Woo, "Scaling Behavior of Sub-Micron MOSFETs on Fully Depleted SOI", Solid-State Electronics, Vol. 39, No. 4, 1996, pp. 445-455.

[2.4] L. Wei, et al., "Double Gate Dynamic Threshold Voltage (DGDT) SOI MOSFETs for Low Power High Performance Designs, *1997 IEEE International SOI Conference Proceedings*, 1997, p. 82.

[2.5] S. Maeda, et al., "Substrate Bias Effect and Source-Drain Breakdown Characteristics on Body-Tied Short-Channel SOI MOSFET's," *IEEE Trans. Electron Devices*, Vol. 46, No. 1, January 1999, pp. 151-158.

[2.6] D. Munteanu, et al., "Generation-Recombination Transient Effects in Partially Depleted SOI Transistors: Systematic Experiments and Simulations," *IEEE Trans. Electron Devices*, Vol. 45, No. 8, Aug. 1998, pp. 1678-1683.

[2.7] A. J. Auerton-Herve, "SOI: Material to Systems", *Proceedings of the 1996 IEDM*, 1996, pp.3-10.