Double-Sided CMOS Fabrication Technology

by

Isaac Lauer
B.S., Electrical Engineering
The Pennsylvania State University, June 1999

Submitted to the Department of Electrical Engineering and Computer Science
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

August 2001
[September 2001]

Author .....
Department of Electrical Engineering and Computer Science
August 28, 2001

Certified by .....
Dimitri A. Antoniadis
Professor
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Double-Sided CMOS Fabrication Technology

by

Isaac Lauer

B.S., Electrical Engineering

The Pennsylvania State University, June 1999

Submitted to the Department of Electrical Engineering and Computer Science on August 28, 2001, in partial fulfillment of the requirements for the degree of Master of Science in Electrical Engineering and Computer Science in Electrical Engineering and Computer Science

Abstract

The focus of this research is the investigation of double-sided CMOS technology. Both Double-Gate (DG) MOSFETs and Double-Sided Interconnects (DSI), collectively referred to as double-sided CMOS, are being examined as extensions of planar CMOS technology with the potential to increase performance and packing density over conventional techniques. The goal of this work is to investigate and develop fabrication technology for double-sided CMOS.

Thesis Supervisor: Dimitri A. Antoniadis
Title: Professor

Acknowledgments

I would first like to thank my advisor, Dimitri Antoniadis, for his guidance and support over the course of this project. I would also like to thank Andy Ritenour for our collaboration on the double-gate portion of this work, and my student mentor Andy Wei for familiarizing me with the fab. Additionally, I'd like to acknowledge the rest of the members and former members of the DAA group whose stay has coincided with mine: Tony Lochtefeld, Keith Jackson, Hasan Nayfeh, Ihsan Djomehri, Duheon Song, Corina Tanassa, Ali Khakifirooz, and Jong Wan Jung. Also, I would like to thank the fellow MTL students outside of my group who have aided my work: Jim Fiorenza, Andy Fan, and John Kymissis, as well as the MTL staff and faculty.

I would also like to acknowledge the SRC, who funded and mentored this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Integrated circuits, especially those based on CMOS (Complementary Metal Oxide Semiconductor) transistors, are the fundamental building block of most electronic equipment at the time of this writing. From digital watches to supercomputers, the MOS transistor is ubiquitous within present-day society. The great success of the semiconductor industry has depended on increasing the performance of electronic devices, mainly through scaling integrated circuits to smaller dimensions. Further scaling is becoming increasingly difficult, making it desirable to examine alternative device and circuit geometries for increased performance and scaling potential.

1.1 Overview of this Work

The focus of this research is the investigation of double-sided CMOS technology. Both Double-Gate (DG) MOSFETs and Double-Sided Interconnects (DSI), collectively referred to as double-sided CMOS, are being examined as extensions of planar CMOS technology with the potential to increase performance and packing density over conventional techniques. The goal of this work is to investigate and develop fabrication technology for double-sided CMOS.

The introductory chapter gives some background material on each area being examined. An overview of current MOSFET technology is given, along with motivation for the study of DG MOSFETs. A brief overview of interconnect technology is then

presented, followed by motivation for double-sided CMOS.

The second chapter contains the details of a self-aligned DG MOSFET fabrication process. Prior work in the area is first presented, along with a summary of the non-idealities associated with that scheme. Another approach is then presented, along with an analysis of the shortcomings of that method. Finally, a new approach is presented and compared to the previous methods.

The third chapter contains a description of a new approach for interconnect technology: double-sided interconnects (DSI). First, some motivation is given, followed by an examination of contamination issues associated with this technology. Then, a process capable of fabricating DSI is presented, as is data obtained from devices fabricated in this manner.

The fourth and final chapter lays out a plan for extending DG MOSFETS and DSI to fully three-dimensional integrated circuits. A process for integrating DG MOSFETS and DSI to form strata used as an individual building block is presented, and a method for integrating different strata into three dimensional circuits is then given.

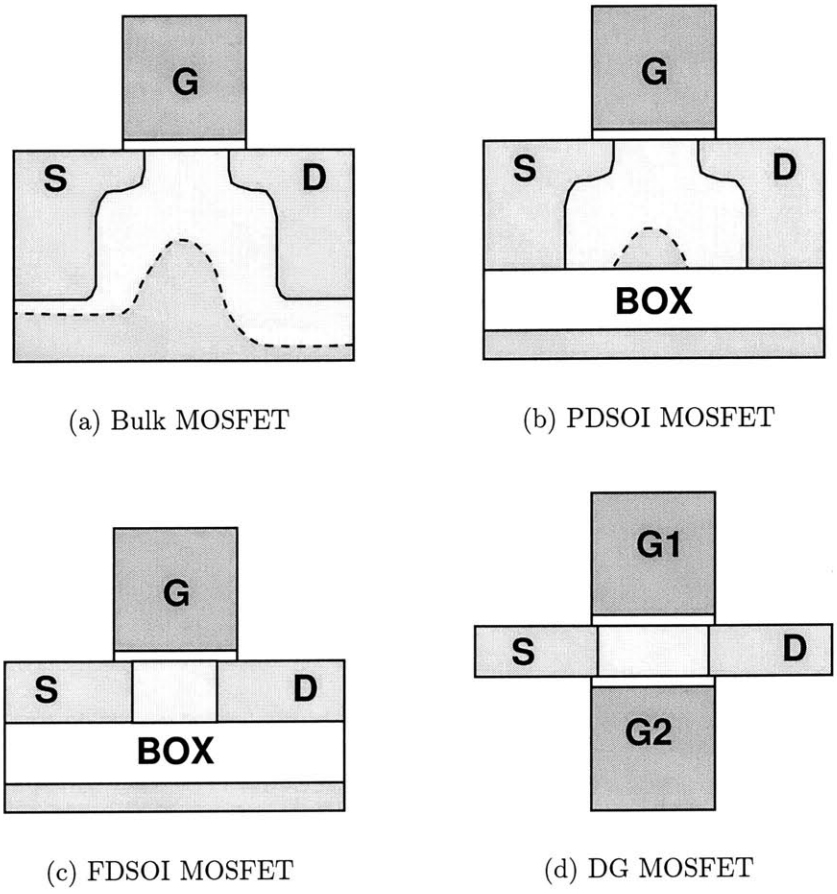
1.2 Double-Gate MOSFETs

DG MOSFETs have been shown to have the potential to be the highest performing MOSFETs for digital applications [1]. A comparison of DG MOSFETs to competing technologies is presented here as background material.

1.2.1 MOSFET Technologies

There are currently four approaches to MOSFET fabrication for logic CMOS. These approaches are: Bulk CMOS, Partially-Depleted (PD)SOI CMOS, Fully-Depleted (FD)SOI CMOS and Double-Gate (DG) CMOS. A schematic cross-section view of each type of device is given in Figure 1-1.

At the time of this writing, bulk CMOS is the incumbent technology in industry. Bulk CMOS has been shown to be highly scalable and has good electrostatic integrity



(a) Bulk MOSFET

(b) PDSOI MOSFET

(c) FDSOI MOSFET

(d) DG MOSFET

Figure 1-1: Logic MOSFET Approaches

at short gate lengths [2]. However, bulk CMOS devices suffer from parasitic junction capacitance between the source/drain regions and the substrate that limits high-speed performance [1].

PDSOI CMOS is currently beginning to transition into commercial products. PDSOI devices are similar to bulk devices except that they are constructed on SOI (silicon on insulator) islands rather than n-type and p-type wells. These SOI islands are thin crystalline silicon films islolated from the suporting bulk silicon substrate by a buried oxide (BOX) layer. PDSOI has the advantage of having reduced parasitic capacitance compared to bulk devices. Also, isolation between the bodies of the device allows circuit designers more freedom. For example, the lack of well implants allows higher packing densities. Scaling PDSOI devices is similar to scaling bulk devices; they both maintain electrostatic integrity down to similar length scales. However,

unlike bulk devices where the body is contacted through the bulk of the wafer, contacting the body of PDSOI devices requires additional layout area and processing steps, so for digital logic the body is typically left uncontacted (floating). While when properly designed the floating state of the body can increase performance in some logic applications, the hysteretic nature of floating body effects make this design difficult [3].

FDSOI CMOS is similar to PDSOI CMOS, except the SOI film thickness is reduced to the point that in saturation, the entire film is depleted. In this case, floating body effects are nearly eliminated [1]. However, FDSOI has poorer scaling potential than bulk because of the lack of screening from the back of the channel [4]. Also, threshold voltage can be strongly dependent on the SOI film thickness which is too thin to be well controlled within the tolerances used to fabricate bulk and PDSOI devices.

DG CMOS has shown the highest potential for scaling. The channel is controlled from either side, leading to increased electrostatic integrity and suppressed short-channel effects. In fact, it has been shown that DG FETs can potentially scale to half of the gate length of bulk CMOS as well as provide twice the drive current at similar lengths [2]. In fact, the lower transverse electric field in the channel further increases drive current by decreasing the degradation of mobility due to scattering at the gate oxide interface [5]. Additionally, SOI film thickness for double-gate can be approximately twice that of FDSOI for the same channel length due to the channel being controlled from either side.

1.2.2 Implementation

There are three possible orientations of a DG MOSFET on a wafer (Figure 1-2). Type I (planar) has the advantage of having the critical silicon channel thickness defined by thin film processing rather than lithography. Other advantages include standard circuit layout and more conventional processing during fabrication. This approach is being examined by IBM [7], and is examined in this work.

Type II allows the gate length to be determined by thin film processing rather

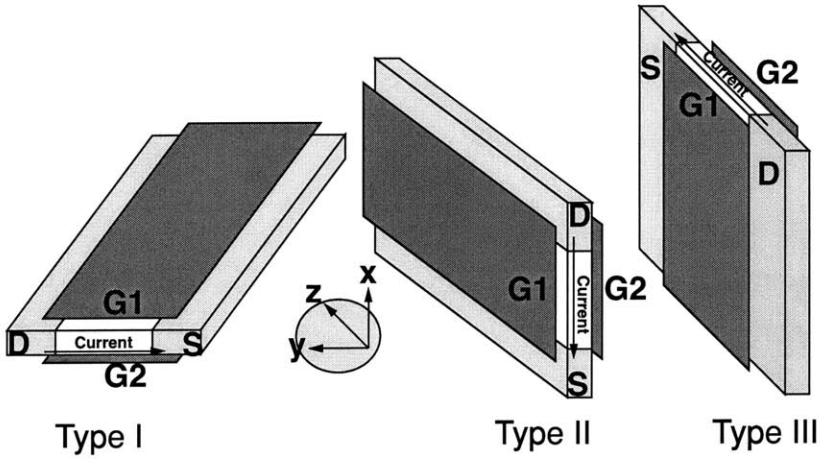


Figure 1-2: DG Approaches (Adapted from Wong et al. [6])

than lithography, which leads to more consistent gate lengths. However, this eliminates the circuit designer's control over gate length and requires different layout than conventional CMOS. An example of research in this area is the Lucent VRG MOSFET [8, 9].

Type III requires gate length to be determined by lithography and fixes the gate width. Circuits requiring varying gate widths must essentially use many devices in parallel rather than wider transistors. An example of research in this area is the Hitachi DELTA FET [10] and the Berkeley FINFET [11, 12, 13].

An additional disadvantage of fully-depleted type II and III devices is that the channel thickness is defined by lithography and must be thinner than the gate length. The gate length is typically the minimum feature size, which defines the most stringent lithography requirements. Type II and III devices thus require more aggressive lithography than type I devices.

1.3 Interconnects

Individual MOSFETs are not very useful until they are connected together into circuits. The wires that connect MOSFETs and other discrete devices together on a chip are referred to as interconnects. While historically circuit performance and relia-

bility has been limited only by device constraints, modern integrated circuits require the parasitics associated with interconnects to be minimized in order to achieve peak performance.

1.3.1 Aluminum

Subtractive Aluminum

The majority of integrated circuits produced at the time of this writing have aluminum interconnects. Aluminum interconnects are typically made using a process known as subtractive aluminum, where a blanket layer of aluminum is deposited and then patterned by a dry plasma etch. Subsequent layers of metal are fabricated in the same way, insulated from each other by deposited interlayer dielectrics. Figure 1-3(a) shows the simplest realization of this structure. After two or three levels of metal have been formed, the topology of the resulting surface eventually prevents additional levels of metal from being added.

CMP

A rather recent innovation is the use of Chemical Mechanical Polishing (CMP) to reduce the topology of the dielectric prior to further metallization. This increase in planarity allows for reduced parasitic capacitance and increase in the number of levels of metal that may be fabricated before the topology precludes additional levels.

Tungsten Plugs

While a purely subtractive aluminum process utilizing CMP can be used for a large number of metal levels, additional planarity and smaller features can be achieved by separating the via-filling process from the interconnect metal deposition. This technique can be achieved by using tungsten damascene plugs as connections from layer to layer and for the contacts to silicon. A simplified damascene process is shown in Figure 1-4. The topology of a subtractive aluminum process utilizing CMP

and tungsten damascene is shown in Figure 1-3(b). Also, tungsten plugs increase interconnect reliability by reducing electromigration.

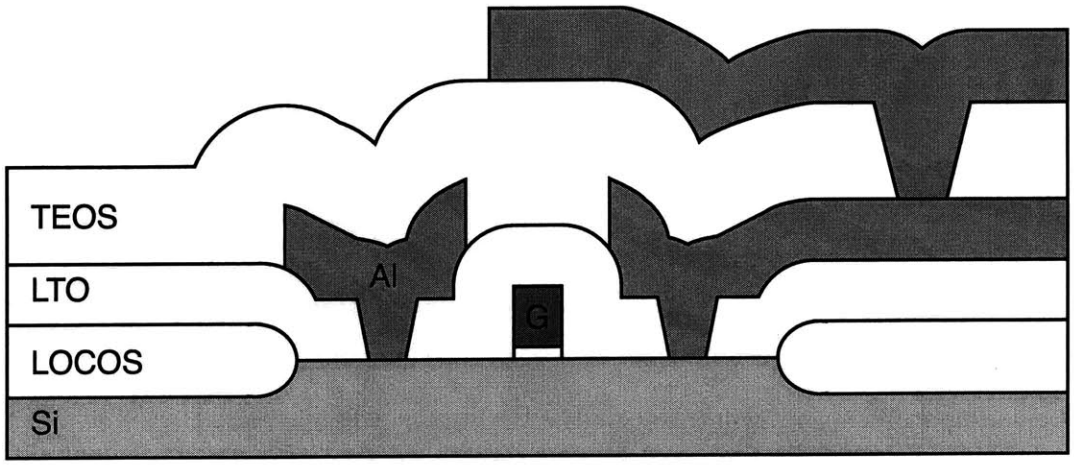
Disadvantages

Aluminum has traditionally been the standard interconnect material for CMOS logic because of its compatibility with CMOS processing. Aluminum has the advantage of being easily patterned by dry etching. Also, when aluminum is introduced into silicon as an impurity, it acts rather benignly as a p-type dopant. However, at small feature sizes, reliability and performance become a concern. At high current densities aluminum suffers from electromigration, a process where the current displaces metal atoms leading to a net flux of atoms through the material, eventually leading to interconnect failure by void formation [14]. Also, the resistance of the lines contributes to RC delays, so any decrease in resistivity of the interconnect material can increase chip performance.

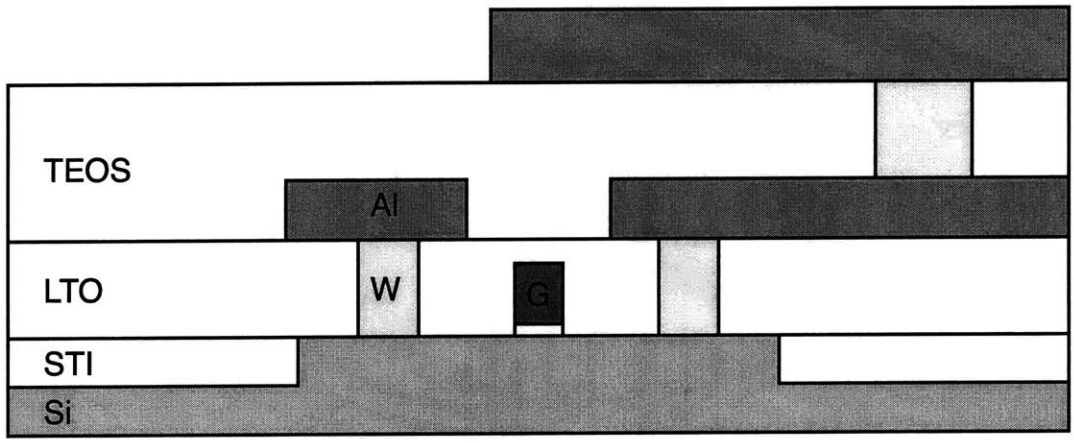
1.3.2 Copper

Because of the limitations associated with aluminum, copper has become the preferred interconnect metal for high-performance processors because of its lower resistivity and higher resistance to electromigration [15]. In order to introduce copper as a successful interconnect material, two advancements had to be made. One advancement was the development of a barrier material that prevents copper from diffusing into the inter-layer dielectric and silicon. The other required advancement was a process capable of patterning copper.

Copper diffuses quickly in silicon dioxide, and more quickly in silicon [15]. Copper forms trap levels near mid-gap in silicon which act as Generation-Recombination (G-R) centers and decrease carrier lifetimes [16]. During normal back-end processing temperatures if no barrier material is employed, the concentration of copper in the underlying silicon layer can become high enough to damage the devices [17, 18]. Additionally, copper diffuses more quickly in the presence of electromagnetic fields [19],

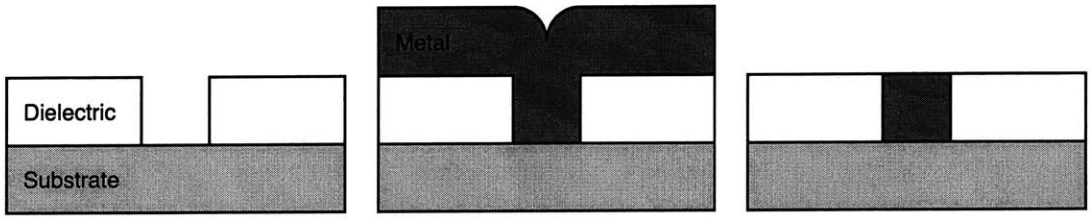


(a) Topology in a subtractive aluminum process



(b) Topology in a subtractive aluminum process utilizing CMP and tungsten damascene plugs

Figure 1-3: Comparison of the Topology of Subtractive Aluminum Processes

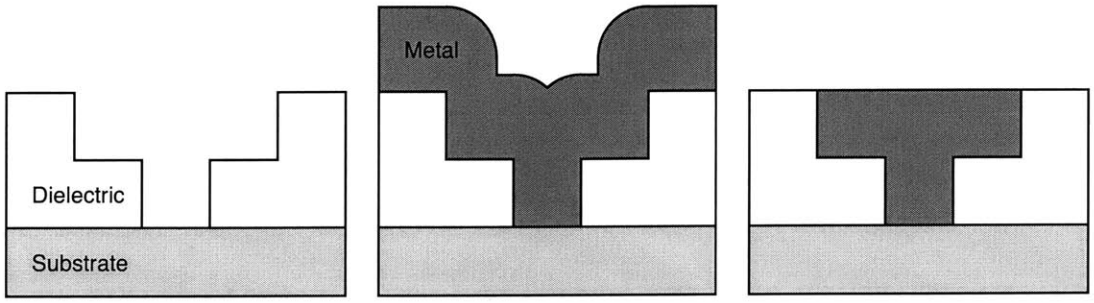


(a) Etch vias in dielectric

(b) Deposit blanket layer of metal

(c) Remove overburden by CMP

Figure 1-4: Damascene Process



(a) Etch lines and vias in dielectric

(b) Deposit blanket layer of metal

(c) Remove overburden by CMP

Figure 1-5: Dual-Damascene Process

so additional contamination of the devices may occur during normal operation of the device. Diffusion barriers were developed to circumvent this problem. When copper is encased in diffusion barrier materials such as tantalum, titanium nitride, etc., copper atoms are unable to diffuse into the silicon device layer [20].

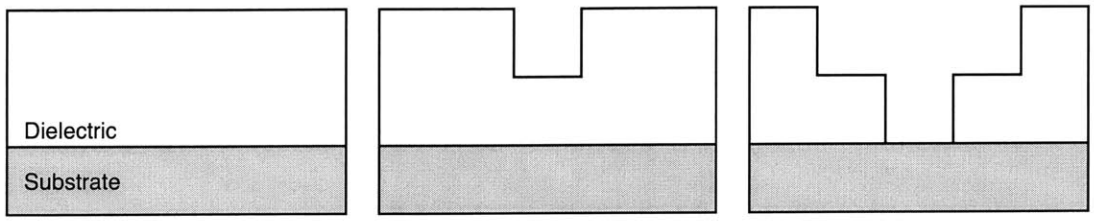
Another difficulty associated with copper metallization is patterning. It is currently impractical to dry etch copper because the operating temperatures required preclude the use of photoresist as the masking layer [15]. Wet etching is also impractical due to the size of the features involved. Because of these difficulties, copper is currently patterned using a dual damascene technique. Dual damascene is a damascene process where both lines and vias are etched and filled at the same time, followed by a single planarization step to remove the overburden [21] (Figure 1-5).

There are three main schemes used to etch the lines and vias. While each scheme has several variants that use different hardmask layers, the idea remains the same. The first technique is a via-first dual damascene etch (Figure 1-6). In a via-first scheme, the through-layer vias are first etched halfway to the underlying layer. Then the interconnect lines are etched, which also completes the via etch. In the line-first technique, the interconnect lines are etched to their final depth first, then the vias are etched (Figure 1-7). Finally, in a so-called "self-aligned" dual-damascene etch, half of the dielectric is deposited, then a hardmask is deposited and patterned, and the final thickness of the dielectric is deposited. Then, both the lines an the vias are

etched in a single step (Figure 1-8).

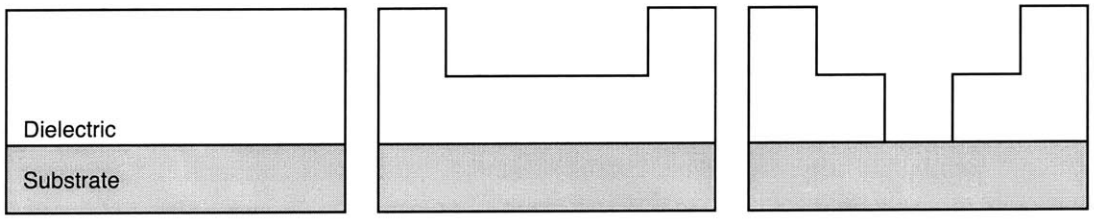
1.4 Double-Sided CMOS

By developing fabrication technology for both double-gate MOSFETs and double-sided interconnects together, there exists the potential to advance CMOS performance by increasing both device performance and interconnect performance with less effort than by examining these processes in isolation. While either technology may be employed separately or not at all, the research is complementary. This work examines a double-gate fabrication approach that employs many of the same techniques as the double-sided interconnect fabrication approach. These same techniques can then be extended to 3D integrated circuits, which may also help advance CMOS performance.



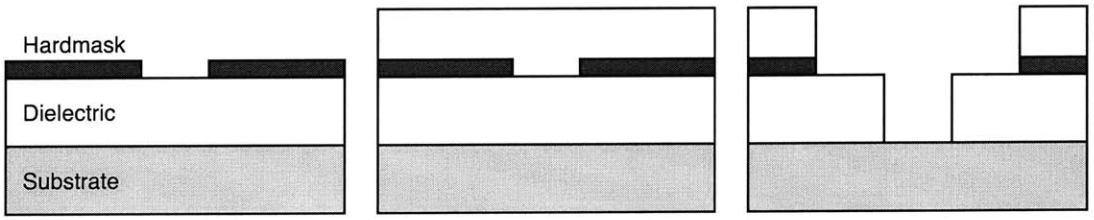
(a) Deposit Dielectric (b) Etch Vias (c) Etch Lines

Figure 1-6: Via-First Dual-Damascene Etch



(a) Deposit Dielectric (b) Etch Lines (c) Etch Vias

Figure 1-7: Line-First Dual-Damascene Etch



(a) Deposit Dielectric and pattern hardmask (b) Deposit Dielectric (c) Etch Lines and Vias

Figure 1-8: Self-Aligned Dual-Damascene Etch

Chapter 2

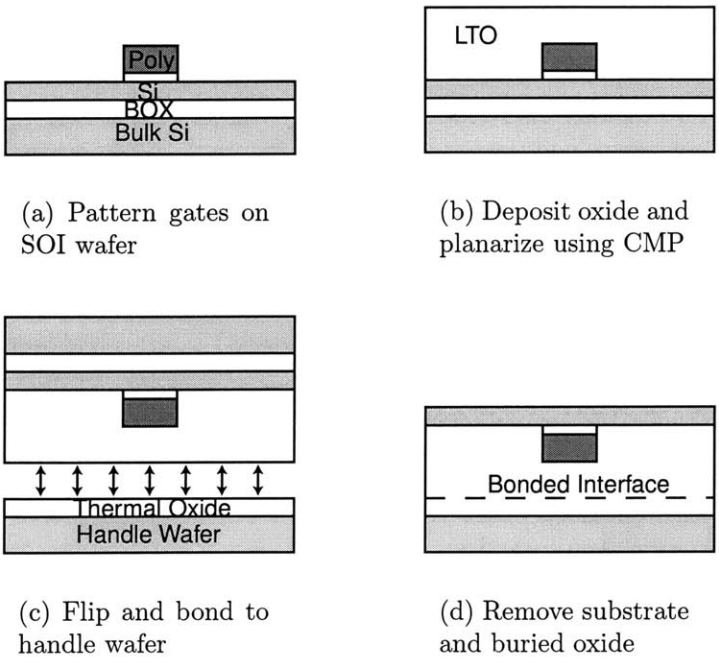
CMBL Alignment

The advantages of the planar DG MOSFET make it a prime candidate for investigation. However, fabrication of these devices, especially alignment of the top and bottom gates, is not trivial.

To maximize current drive and minimize parasitic capacitances in double-gate devices, the top and bottom gates should have near perfect alignment [22]. Alignment to within very high tolerances is currently only practically achieved through self-alignment. Developing a manufacturable self-aligned double-gate process that can be scaled to dimensions shorter than is possible for bulk silicon remains a key challenge.

2.1 Flip, Bond, and Transfer

The first planar DG MOSFETs were produced by Tanaka et al. in 1991 [23] using a flip, bond, and transfer technique (Figure 2-1) to bury the bottom gate. After the step shown in Figure 2-1(d), another gate oxide was grown, polysilicon was deposited and patterned, and the source and drain were implanted to complete the device. However, alignment between the top and bottom gates was performed optically by photolithography. In order to align the top gate to the bottom gate, an oversized bottom gate was employed. While easing alignment, the oversized bottom gate introduces parasitic capacitances that severely limit the potential of these devices.



(a) Pattern gates on SOI wafer

(b) Deposit oxide and planarize using CMP

(c) Flip and bond to handle wafer

(d) Remove substrate and buried oxide

Figure 2-1: Flip, Bond, and Transfer Process

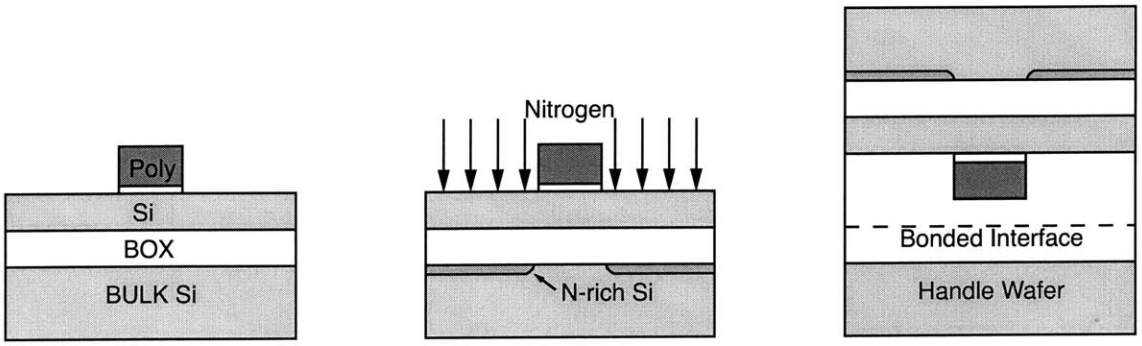
2.2 CMBL

In order to take advantage of the flip, bond, and transfer process and allow self-alignment of the top and bottom gates, a technique known as Chemical Modification of Buried Layers (CMBL) was developed by Ritenour of MIT [24].

CMBL takes advantage of the differential etch rate of nitrogen-rich silicon compared to that of untreated silicon. Silicon that has been implanted with a sufficiently high dose of nitrogen has up to a 100x reduced etch rate in an aqueous solution of tetramethyl ammonium hydroxide (TMAH) compared to non-implanted silicon. By using existing polysilicon gates as an implant mask, a self-aligned nitrogen-rich silicon region can be introduced into a device, which can then later be used to fabricate a self-aligned gate.

2.3 Original Negative Image Technique

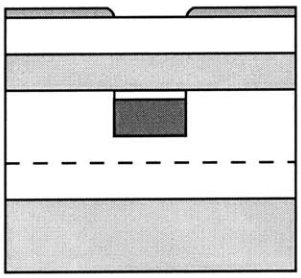
An outline of the original process developed by Ritenour to utilize CMBL for the fabrication of self-aligned double gates is shown in Figure 2-2.



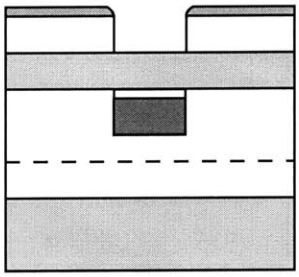
(a) Pattern gates on SOI wafer

(b) Implant nitrogen

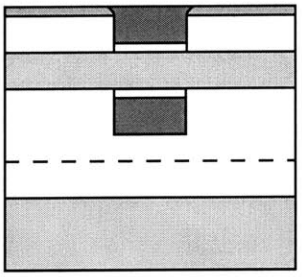
(c) Deposit LTO, planarize, flip and bond



(d) Etch back substrate



(e) Etch self-aligned trench



(f) Grow gate oxide, deposit poly, planarize

Figure 2-2: Original Negative Image CMBL Process

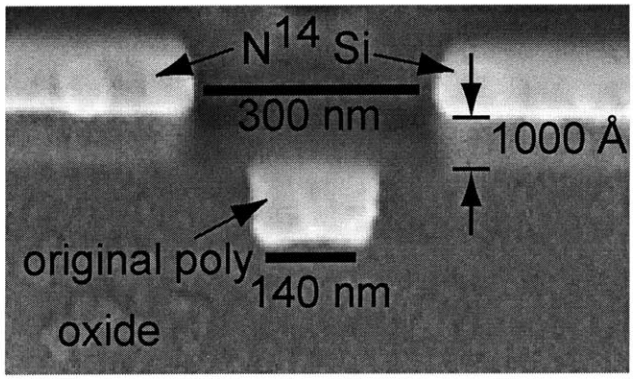


Figure 2-3: Self-Aligned Hardmask and Trench Fabricated Using CMBL

First, polysilicon gates are patterned on an SOI substrate (Figure 2-2(a)). Then, a $10^{17}cm^{-2}$ nitrogen implant defines self-aligned nitrogen-rich silicon regions in the substrate with a 100:1 etch selectivity relative to silicon in TMAH after annealing (Figure 2-2(b)). A layer of LPCVD low temperature oxide (LTO) is then deposited. This LTO layer is then planarized by CMP as preparation for wafer-bonding. The device wafer is then thermally bonded to a handle wafer (Figure 2-2(c)). The original silicon substrate is then mechanically ground away, leaving around $10\mu m$ of silicon. The remaining silicon substrate is then chemically etched away in TMAH, leaving the nitrogen-rich silicon behind (Figure 2-2(d)). This nitrogen-rich silicon is then used as a hardmask to etch a trench in the buried oxide that is self-aligned to the original gate (Figure 2-2(e)). A new gate oxide is grown in the trench, then the trench is filled with polysilicon. The polysilicon overburden can then be removed by CMP, leaving a self-aligned gate (Figure 2-2(f)).

Figure 2-3 [24] shows a nitrogen-rich silicon hardmask and a trench self-aligned to an original polysilicon gate. In this case, the self-aligned trench is over-sized compared to the original gate. This should be controllable by altering the implant dose and the anneal conditions.

While it is possible to create self-aligned gates with this technique, several factors make this approach undesirable. First, the mechanical grinding of the substrate must be very uniform in order to leave a thin layer of the substrate over an entire wafer. Second, in order to have the selectivity necessary to leave the self-aligned nitrogen-

rich silicon regions intact after the chemical etch-back, a nitrogen dose on the order of $10^{17}cm^{-2}$ is required. Implanting this many ions into a device will certainly damage it. Finally, it is difficult to polish polysilicon using CMP, making the removal of the overburden difficult.

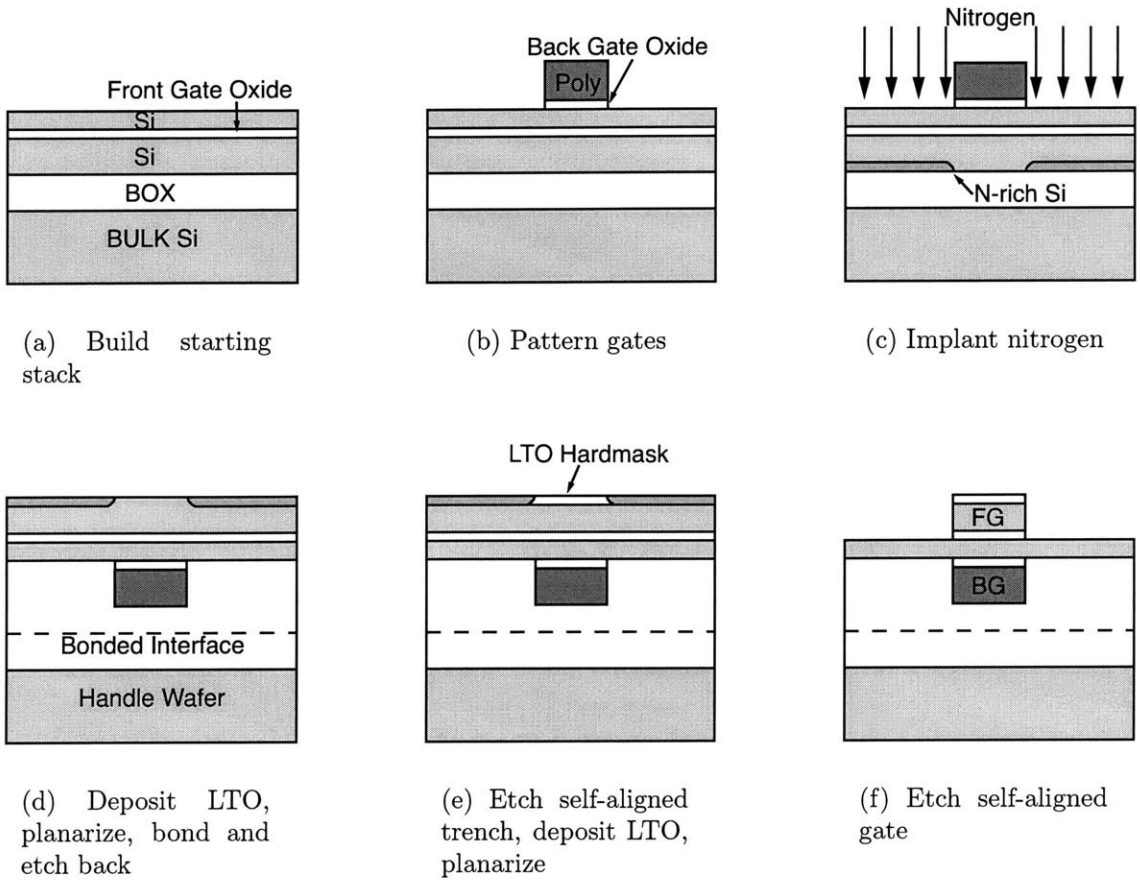
2.4 Improved Negative-Image Technique

A new technique was developed in collaboration with Ritenour to take advantage of the CMBL technique, while attempting to reduce the shortcomings of the first approach. By using an imaging layer, the precise mechanical thinning, high required selectivity, and polysilicon CMP could be done away with.

2.4.1 Process Flow

The improved process is shown in Figure 2-4. First a starting stack is formed on an SOI substrate (Figure 2-4(a)). Then, a gate stack is added and patterned (Figure 2-4(b)). Using this gate as a mask, a $3\times10^{16}cm^{-2}$ nitrogen implant which gives a 10:1 etch selectivity in TMAH is used to form a nitrogen-rich silicon region in the thin silicon imaging layer/back gate material (Figure 2-4(c)). After forming the nitrogen-rich silicon region, LTO is deposited and planarized so that the wafer can be bonded to a handle wafer. The silicon substrate is then removed (Figure 2-4(d)). The exposed SOI layer is then chemically etched using a timed TMAH dip. The TMAH etch leaves behind a trench aligned to the original gate. After formation of the self-aligned trench, the image is reversed by depositing and planarizing LTO, leaving an oxide hard mask (Figure 2-4(e)). The oxide hard mask is then used to etch a self-aligned gate (Figure 2-4(f)).

A demonstration structure was fabricated using this technique as a proof-of-concept. This demonstration structure used an LTO layer (Figure 2-5) rather than the starting stack shown in Figure 2-4(a) because of the time and expense involved with building the starting stack. An SEM image of the resulting double-gate structure is shown in Figure 2-6.



(a) Build starting stack

(b) Pattern gates

(c) Implant nitrogen

(d) Deposit LTO, planarize, bond and etch back

(e) Etch self-aligned trench, deposit LTO, planarize

(f) Etch self-aligned gate

Figure 2-4: Improved Negative Image CMBL Process

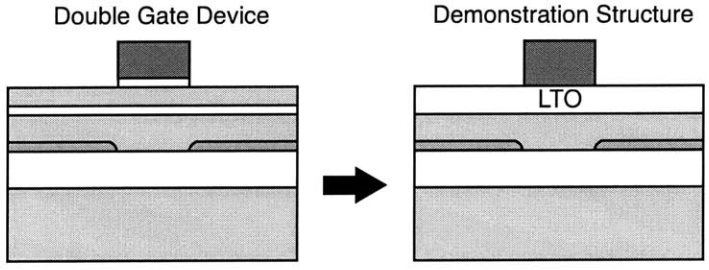


Figure 2-5: Demonstration Structure

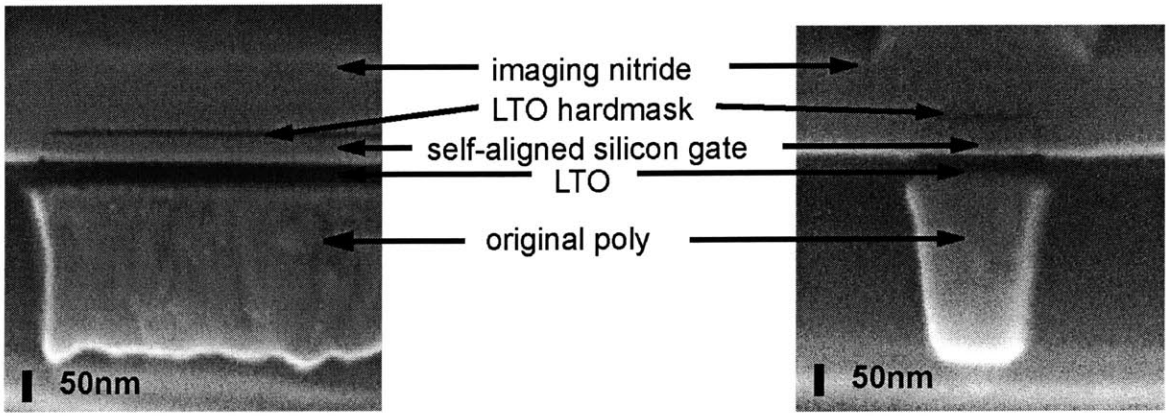


Figure 2-6: SEM of Self-Aligned Gates

Non-implanted Devices	Implanted Devices
Form gates	Form gates
Implant gate	Implant gate
Deposit hard mask	Deposit hard mask
Etch gate	Etch gate
	Nitrogen implant
Reoxidation	Reoxidation
S/D Implants	S/D Implants
RTA	RTA

Table 2.1: Abbreviated Process Flow for Determining the Electrical Effects of a High-Dose Nitrogen Implant

While this technique can be used to produce self-aligned gates, there are still drawbacks that make this approach undesirable. One problem is the complicated starting stack, which requires the use of a bonded gate oxide. Another problem is that the nitrogen dose is still relatively high, potentially damaging the devices.

2.4.2 Test Devices

In order to determine the electrical effects of the high-dose nitrogen implant, NMOS devices were fabricated with and without a nitrogen implant. An abbreviated process flow for this experiment is given in Table 2.1.

By measuring the electrical characteristics of the implanted and non-implanted devices and making comparisons between the two, some empirical conclusions can be made about the effects of the implant. The log scale $I_D - V_G$ plot shown in Figure 2-

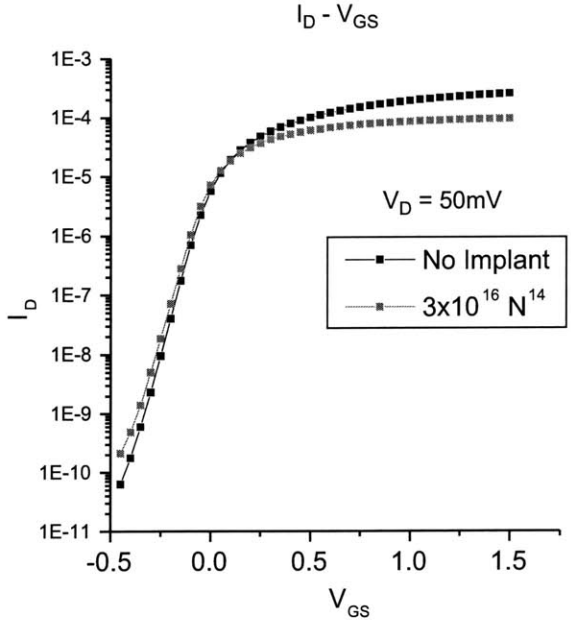


Figure 2-7: Comparison of $I_D - V_G$ Between Implanted and Non-Implanted Devices

7 can be used to summarize the detrimental effects. First, there is an appreciable loss in drive current. Also, there is an increase in inverse subthreshold slope (S). These problems can be examined in more detail by looking at additional MOSFET characteristics.

The capacitance vs. voltage (C-V) plot shown in Figure 2-8 indicates increased poly-depletion for the nitrogen implanted devices. The gate oxide capacitance (C_{ox}) is reduced, and at high fields it is reduced further, clearly indicating poly-depletion.

Using the linear $I_D - V_G$ plots for a family of devices shown in Figure 2-9, the source/drain series resistance can be extracted using the Terada-Muta method [25]. For these devices, the nitrogen implant increased the series resistance from 70Ω to 245Ω.

The reduced drive current, increased source drain resistance, and reduced gate capacitance can all be explained by lower doping in the source, drain and gate regions. This lower doping is most likely the result of reduced dopant activation due to the high nitrogen concentration in these regions. Reduced carrier mobility is also likely evident.

The resistivity of uniformly doped n-type silicon is given by Eq. (2.1)[26]. To the

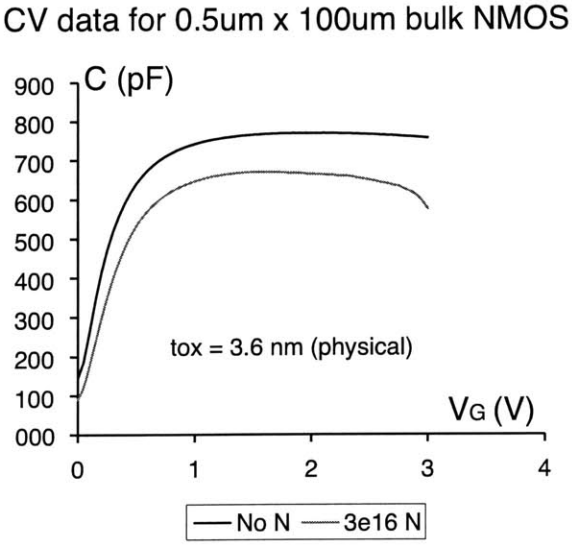


Figure 2-8: Comparison of C-V Characteristics Between Implanted and Non-Implanted Devices

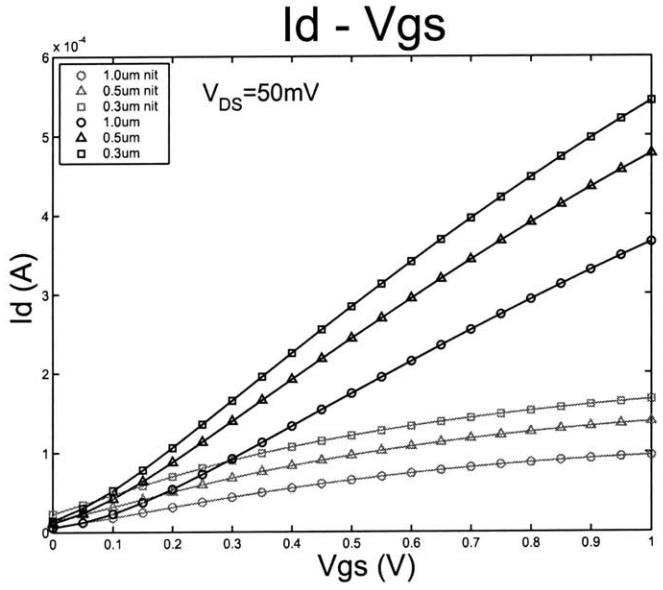


Figure 2-9: Linear Plot of I_D-V_{GS} for Implanted and Non-Implanted Devices with Varying Gate Lengths used for Series Resistance Extraction

first order, this means that the increased series resistance can be explained by reduced doping, although reduced mobility could be caused by the nitrogen and also likely increases the series resistance. Recent work at Lucent has indicated that nitrogen reduces mobility by acting as scattering sites by disrupting the silicon lattice and forming interstitials [27]. The 3.5× increase in series resistance is probably due to a combination of reduced doping and deceased electron mobility.

$$\rho = \frac{1}{q\mu_n N_D} \tag{2.1}$$

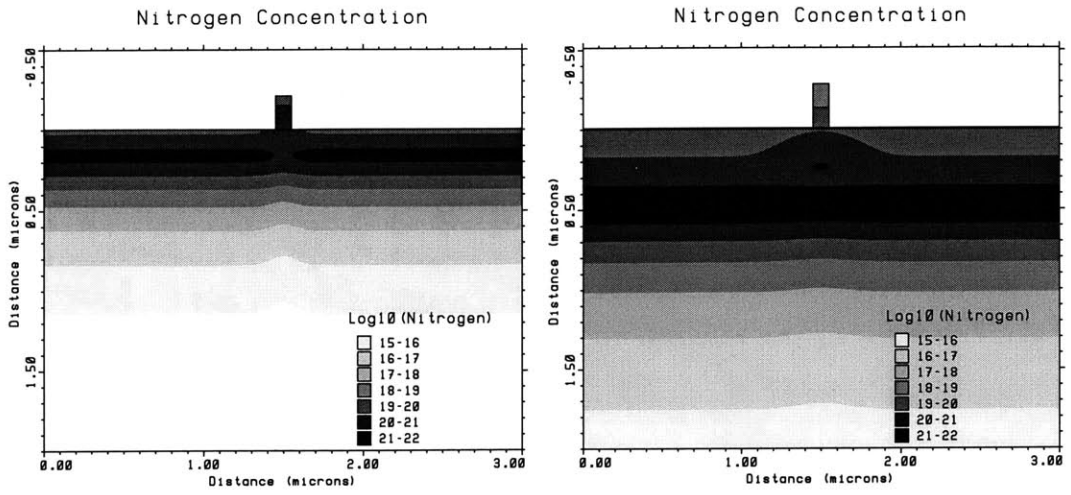
The decreased gate capacitance can also be explained by reduced doping. In a MOSFET with polysilicon gates, the gate capacitance is affected by the depletion of the gate. This effect is known as poly-depletion. The gate capacitance C_{max} can be considered as the gate capacitance without poly-depletion C_{ox} in series with the gate depletion capacitance, as shown in Eq. (2.2)[28]. The last term in Eq. (2.2) shows the dependence on gate doping.

$$\frac{1}{C_{max}} = \frac{1}{C_{ox}} + \sqrt{\frac{8kT}{\epsilon_{Si} q^2 N_D}} \tag{2.2}$$

The increased poly-depletion also explains the increased inverse subthreshold slope, which is given by Eq. (2.3) [28], where C_{dm} is bulk depletion capacitance and C_{max} is the gate capacitance including poly-depletion.

$$S = 2.3\frac{kT}{q}\left(1 + \frac{C_{dm}}{C_{max}}\right) \tag{2.3}$$

The combination of increased series resistance and decreased gate capacitance can be explained by reduced dopant activation caused by the nitrogen implant. If high-performance DG MOSFETS are to be produced using a CMBL method, it is clear that the device must somehow be protected from the effects of the nitrogen implant.



(a) Negative Image Approach

(b) Positive Image Approach

Figure 2-10: Comparison of Nitrogen Concentration in Devices for Negative Image and Positive Image Approaches

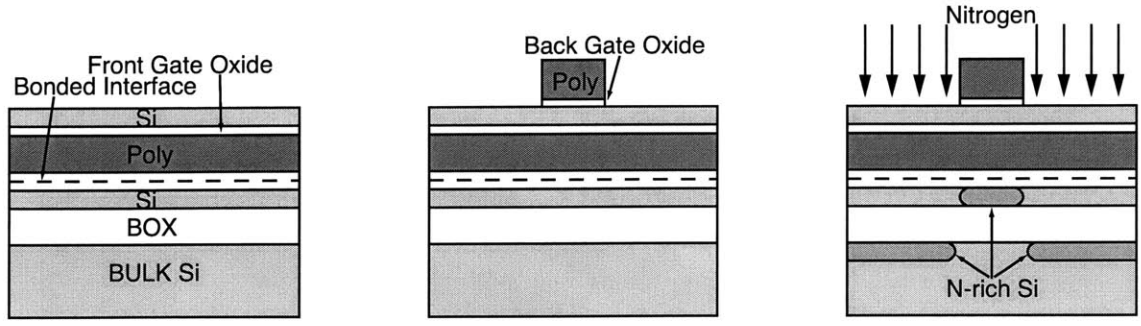
2.5 Positive Image Method

A new technique was developed as another attempt to utilize CMBL to pattern self-aligned gates. Because a high-concentration nitrogen region is essential for achieving the etch-rate selectivity required for patterning by CMBL, lower doses are impractical for implants deep enough to image a gate. Even though the implant dose cannot be reduced, the implant can be moved deeper into the device, leaving a lower concentration of nitrogen in the device itself by implanting the whole way through it.

An analytical comparison of the nitrogen concentration found in devices using the negative and positive image techniques is shown in Figure 2-10. In the positive image approach, the nitrogen concentration in the device is reduced by at least an order of magnitude.

2.5.1 Process Flow

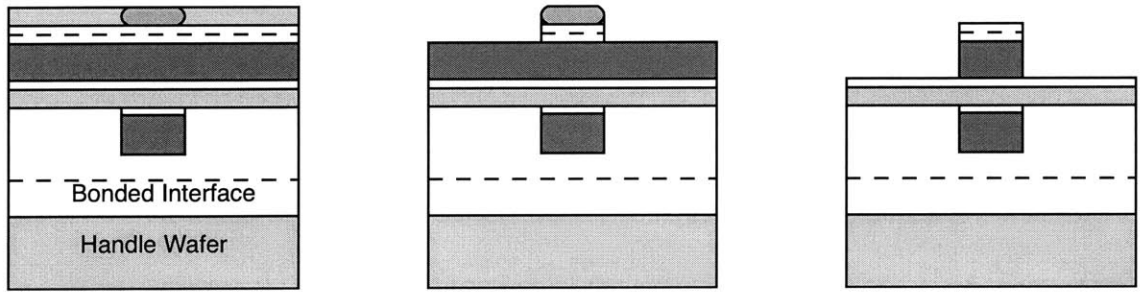
Figure 2-11 gives an overview of a process that utilizes a higher-energy implant. First, a starting stack is built using a flip, bond, and transfer technique (Figure 2-11(a)). Then, gates are patterned on this stack (Figure 2-11(b)). Nitrogen is then



(a) Build starting stack

(b) Pattern gates

(c) Implant nitrogen



(d) Deposit LTO, planarize, bond and etch back

(e) Etch self-aligned trench, deposit LTO, planarize

(f) Etch self-aligned gate

Figure 2-11: Positive Image CMBL Process

implanted at high energy into the device. Because of the additional thickness of the gate above the surrounding substrate, a self-aligned nitrogen-rich silicon region is formed directly under the gate (Figure 2-11(c)). The device film is then transferred to a handle wafer(Figure 2-11(d)). The silicon imaging layer can then be etched in TMAH to leave a self-aligned hardmask. This hardmask can then be used to etch the oxide layer underneath (Figure 2-11(e)). This oxide hardmask can then be used to etch the self-aligned gate(Figure 2-11(f)).

2.5.2 Test Devices

Because this technique requires implanting high doses of nitrogen through the gate oxide, test devices were constructed to determine the electrical effects of the high-

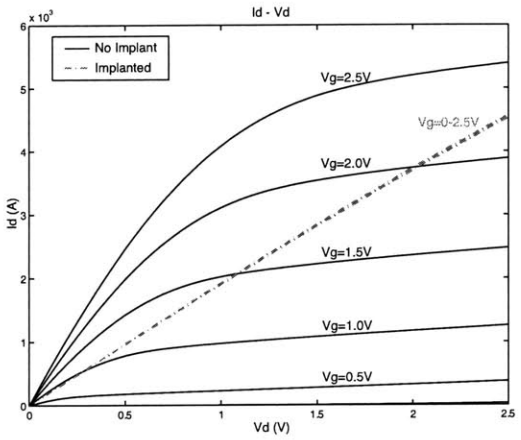


Figure 2-12: 1 μm $\times$ 25 μm Bulk NMOSFETs with and without Nitrogen Implant

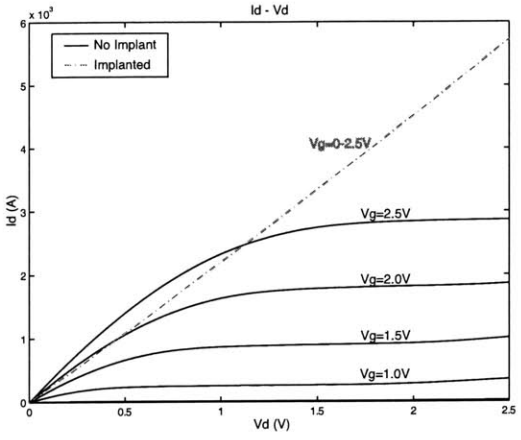


Figure 2-13: 1 μm $\times$ 25 μm SOI NMOSFETs with and without Nitrogen Implant

energy nitrogen implant.

NMOS devices were fabricated on bulk and SOI wafers. Nitrogen was implanted into every other die on the wafer in a checkerboard pattern by using photoresist to block the implant. By comparing devices in adjacent dies, it is possible to conclusively determine that any differences are due to the implant itself and not any processing errors. Plots of $I_D - V_D$ for bulk devices with and without the implant are given in Figure 2-12 and plots for the SOI devices are given in Figure 2-13. Clearly, the nitrogen implant has completely destroyed the devices. Further analysis reveals that the source/drain regions are shorted. A possible explanation is that the high energy implant amorphized the channel. During the post-implant anneal, the channel would recrystallize into poly-silicon because of the lack of a good template, which would cause the source/drain dopants to diffuse across the channel and short together.

2.6 Scalability

While none of the CMBL techniques examined thus far appear very promising, additional advancements or ideas may make it desirable to re-examine this approach. With this in mind, the fundamental scalability limits of the implantation technique for imaging must be taken into account. Analytical and Monte-Carlo (as verifica-

tion) simulations have been performed in order to determine whether either the first two processes presented (negative image) or the final process (positive image) can be scaled to dimensions where DG devices are necessary. Two scalability issues were examined. The first issue is minimum feature size, which should be less than 50nm to be of interest. The second issue is feature pitch, which should be around twice the minimum feature size. It should be noted that these simulations do not necessarily represent hard limits, rather, they show the relative scalability of the negative and positive image techniques.

An assumption made during these simulations was that at least a factor of two modulation in nitrogen concentration is necessary to image any feature. This assumption is most likely rather optimistic, but was chosen because no experimental data was available for the minimum required modulation. Another assumption is that once a feature is imaged, other techniques can be used to modify the final feature size if the image is over-sized or under-sized. Only self-registration is required.

2.6.1 Negative Image

In a negative image scheme, the gate is imaged using a region of non-implanted silicon surrounded by nitrogen implanted silicon. For the techniques presented in the previous sections, this has the advantage of being a lower energy implant than the positive image scheme. This means that the lateral straggle of the implant is reduced.

A plot of peak nitrogen contours for varying gate lengths for an isolated gate is shown in Figure 2-14. For this example, the minimum imageable feature size is 90nm. A plot of varying gate pitches for 90nm gates is given in Figure 2-15. For this example, the minimum feature pitch is 300nm.

2.6.2 Positive Image

In a positive imaging scheme, the gate is imaged directly using a self-aligned nitrogen-rich silicon region. This has processing advantages in that it does not require some type of image reversal, but it means that in the given process, the implant has to be

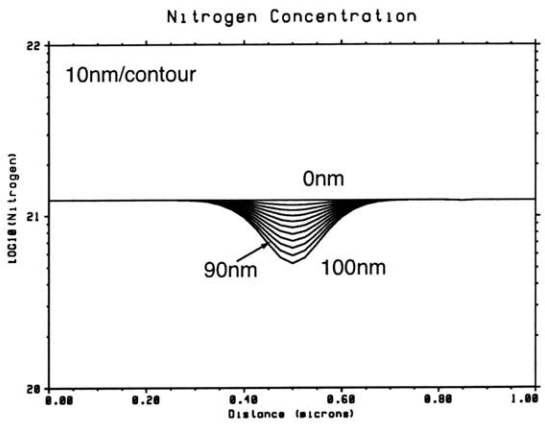


Figure 2-14: Nitrogen Concentration Profiles for Various Gate Lengths using a Negative Image Approach

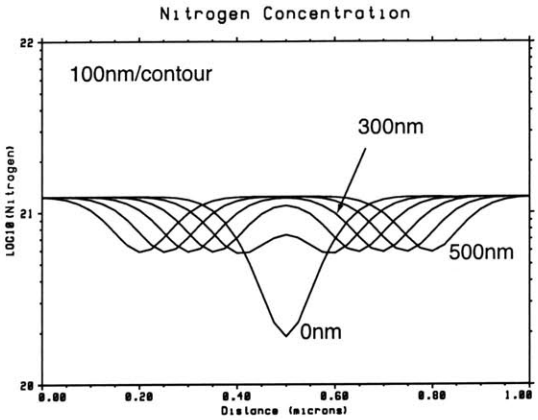


Figure 2-15: Nitrogen Concentration Profiles for Various Gate Pitches using a Negative Image Approach($L_{gate} = 90nm$)

done at high energy, leading to a large lateral straggle.

A plot of peak nitrogen contours for varying gate lengths for an isolated gate is shown in Figure 2-16. For this example, the minimum imageable feature size is 10nm. A plot of varying gate pitches for 10nm gates is given in Figure 2-17. For this example, the minimum feature pitch is 800nm. For comparison with the feature pitch of the negative image approach, the nitrogen concentration profiles for varying gate pitches for 90nm gates is given in Figure 2-18, where the minimum feature pitch is 400nm.

2.6.3 Scaling Conclusions

As can be seen by comparing Figure 2-14 with Figure 2-16, the positive image technique can be used to image smaller gates, and can give higher contrast at the same gate length. However, by comparing Figure 2-15 with Figure 2-18, the minimum gate pitch associated with the positive technique is larger due to the increased straggle associated with the high-energy implant. It should be noted while only the positive image technique can image small enough feature sizes to be of interest in double-gate work, neither technique can be used to produce desirable gate pitches.

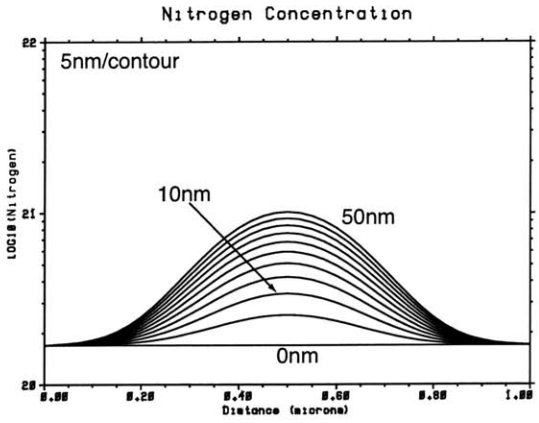


Figure 2-16: Nitrogen Concentration Profiles for Various Gate Lengths using a Positive Image Approach

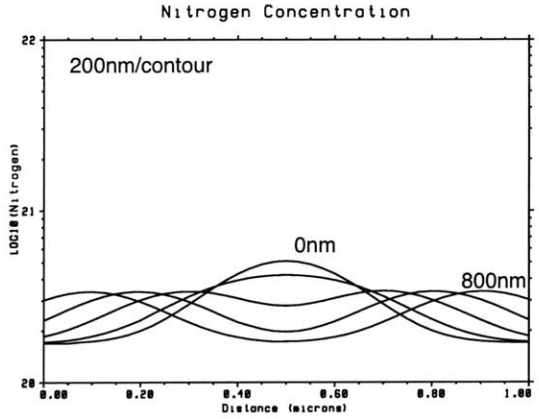


Figure 2-17: Nitrogen Concentration Profiles for Various Gate Pitches using a Positive Image Approach ($L_{gate} = 10nm$)

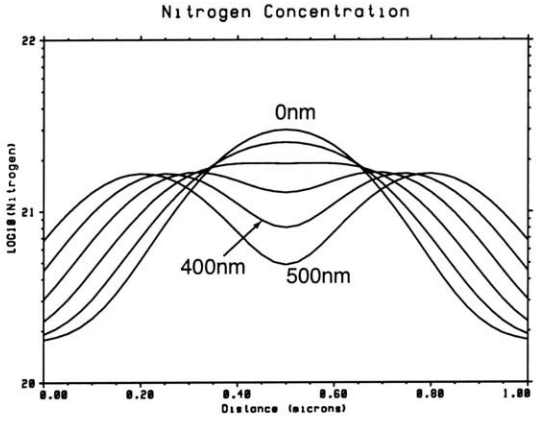


Figure 2-18: Nitrogen Concentration Profiles for Various Gate Pitches using a Positive Image Approach($L_{gate} = 90nm$)

2.7 Conclusions

To this point, no CMBL process has been developed that can produce high-performance devices. While it is possible to utilize CMBL in several different ways to image self-aligned gates, the high nitrogen dose required always damages the device. Even if a process can be developed that either shields the device from the implant or anneals out the damage, it appears that this technique is still not desirable because of scaling issues. The fundamental property of implant straggle makes imaging close-packed gates impossible.

Chapter 3

Double-Sided Interconnects

While designing processes that bury the back-gate poly in DG MOSFETs, it became apparent that it is possible to bury more than just the back gates. In fact, it should be possible to build interconnect levels before bonding, which results in buried interconnects after bonding. By building additional layers of interconnects after the bond and transfer, double-sided interconnects (DSI) can be built. DSI can potentially improve interconnect routing, improving performance, and could possibly allow for more levels of interconnects. First some motivation for double-sided interconnects is given, followed by an analysis of contamination issues and process capable of fabricating a proof-of-concept.

3.1 Motivation for Double-Sided Interconnects

While advancements such as copper interconnects can help improve performance, further improvement can come from improved routing schemes. While lower resistivity materials such as copper can improve RC delays, shorter interconnects have both lower resistance and capacitance for the same material. DSI can allow for alternative routing schemes, and possibly allow more levels of metal to improve conventional routing, by allowing interconnects on both sides of an SOI film. In a typical logic design today, each level blocks around 12%-15% of the wiring level underneath it [1]. This suggests that the useful number of interconnect layers is limited to around seven.

While more levels of interconnects are already planned, there are diminishing returns on the investment for more levels. A double-sided approach has the potential to widen this bottleneck.

Another advantage of DSI is the extensibility to 3D. Once the techniques necessary to fabricate double-sided interconnects are developed, a simple extension of them can be used to fabricate high-density 3D circuits. The extension to 3D is covered in Chapter 4.

3.2 Contamination Issues

This DSI process brings up some issues regarding contamination of both the device wafer and the equipment that is used to process it. For planarity, the buried interconnects should be patterned using a damascene technique. Damascene CMP is usually considered the "dirtiest" back-end process. Back-end processes are not kept under the stringent contamination controls that are associated with front end processing. However, in a DSI process, the delicate silicon layer is exposed and undergoes additional front-end processing after it has already been exposed to the contaminants associated with back-end processing. Experiments were performed in order to measure the level of contaminants the wafer is exposed to during metal CMP. Also, the potential effects of this contamination on the exposed wafers as well as cross-contamination risks were examined.

3.2.1 Contaminant Exposure

The first experiment was used to determine the level of contaminants that the wafer was exposed to during metal CMP. The wafers were processed normally, then contamination levels were examined with no cleans performed.

Every processing step (except perhaps cleans) introduces some level of new contaminants onto the wafer surface. To determine the severity of contamination from metal CMP, a control group and a comparison group were used. The control was a new wafer that had been through an RCA clean. This is considered the cleanest a

wafer can be. The comparison group was a wafer that had been exposed to dielectric CMP. This level of cleanliness is also considered acceptable because dielectric CMP is used prior to additional front-end processing without concern for contamination.

While the interconnect material used in subsequent DSI experiments is tungsten, the CMP machine used for this experiment is also used for copper damascene. As mentioned earlier, copper forms deep traps in silicon that reduce carrier lifetimes. Any copper exposure the wafers endure will surely cause device damage. To reduce the risk of copper exposure, the machine was brought into a "clean" state. During CMP, the wafers only make contact with the insert pad in the wafer carrier, the polishing pad, and the slurry. Additionally, the pad is in contact with a pad-conditioner that could be a source of contamination. Because the materials the wafer makes contact with are consumables, the machine appears "clean" when the carrier insert, polishing pad, and pad conditioner are replaced.

The experimental procedure consisted of the following procedure. New silicon wafers were cleaned using a standard RCA clean. One of these wafers were kept as a control (CTL). Then, the other wafers were processed. One of the wafers was polished for two minutes on the dielectric CMP machine (CMP1), while another was polished for two minutes on the metal CMP machine after it was brought into the "clean" state (CMP2). Then, all of the wafers were analyzed using Total Reflection X-Ray Fluorescence (TXRF).

The TXRF results are summarized in Table 3.1, where uncertainty values ($\pm$) are one standard deviation calculated from instrument reproducibility and background signal-to-noise ratios. Concentrations less than the detection limits of the machine are noted with a "$<$" sign.

The concentration of copper ions on wafer CMP2 is comparable to that of CMP1, meaning that the procedure used to clean the machine was successful. However, the level of other metals is much higher on wafer CMP2. The metals appear to be coming from the tungsten polishing slurry used for this experiment. An analysis of the metals in the slurry is given in Table 3.2. It is evident that the slurry contains high levels of the metals detected by TXRF.

Test	Ca	Ti	Cr	Mn	Fe	Ni	Cu	Zn
CTL(1)	< 6	< 4	< 1.0	< 0.8	< 0.6	< 0.5	< 0.5	1.8 ± 0.3
CTL(2)	< 6	< 2	1.7 ± 0.6	< 0.9	13 ± 1	1.2 ± 0.3	< 0.5	0.9 ± 0.3
CTL(3)	< 6	< 4	< 1.0	< 0.8	< 0.7	< 0.5	< 0.5	1.9 ± 0.3
CMP1(1)	22 ± 4	< 2	< 1.2	< 1.0	214 ± 13	1.9 ± 0.4	4.5 ± 0.5	8 ± 0.6
CMP1(2)	11 ± 3	< 2	< 1.2	< 1.0	39 ± 3	5.8 ± 0.6	2.9 ± 0.4	3.2 ± 0.5
CMP1(3)	9 ± 3	< 4	< 1.0	< 0.8	34 ± 2	10 ± 0.8	9.3 ± 0.7	3.7 ± 0.4
CMP2(1)	< 20	2600 ± 16	92 ± 6	18 ± 2	180 ± 11	80 ± 5	15.7 ± 1.3	64 ± 4
CMP2(2)	< 16	2200 ± 130	59 ± 4	9.5 ± 1.5	176 ± 11	69 ± 4	10.3 ± 0.9	49 ± 3
CMP2(3)	130 ± 11	2150 ± 130	64 ± 4	10.3 ± 1.4	163 ± 10	65 ± 4	< 2.5	38 ± 2

Table 3.1: TXRF Results in Units of 10^{10} atoms/cm^2

Element	Concentration (ppm)
Ca	5.8
Ti	90
Cr	1.3
Mn	1.0
Fe	12
Ni	3.6
Cu	1.0
Zn	4.0

Table 3.2: Metal Ion Concentration in Tungsten Polishing Slurry

Split A	Split B	Split C	Split D
RCA	RCA	RCA	RCA
grow 1K oxide	grow 1K oxide	grow 1K oxide	grow 1K oxide
		polish 1min dielectric CMP	polish 1min metal CMP
		water polish	water polish
		post-CMP scrub	
		double piranha	ultrasonic HCl
	anneal 850°C 1hr	anneal 850°C 1hr	anneal 850°C 1hr
measure	measure	measure	measure

Table 3.3: Post-Contamination Experimental Procedures

3.2.2 Post-Contamination Effects

The contaminant exposure experiment revealed that wafers undergoing metal CMP are in fact exposed to more contaminants than wafers that go through a normal front-end process. The impact of this exposure was then assessed in another set of experiments.

To determine the effects of the exposure to contamination, an experiment was devised comparing control wafers, wafers exposed to dielectric CMP, and wafers exposed to metal CMP. The processing done to these wafers is summarized in Table 3.3.

Splits A and B were the control groups. Split A is an ideal case, where new wafers are RCA cleaned and a 1000Å dry oxide is grown in a gate oxidation tube. Split B is used to measure the impact of the contaminant drive-in anneal on a clean wafer. Split C consists of wafers that get exposed to dielectric CMP, while split D is made up of wafers that are exposed to metal CMP.

These wafers were measured using a Surface Charge Analyzer (SCA). Among other

parameters, this device is capable of measuring fixed oxide charge (Q_{ox}), interface trap states (D_{it}), and carrier lifetimes (T_s). A summary of the results obtained is given in Figure 3-1. Each bar represents the average of five measurements each on three wafers for a total of 15 measurements. The spread bar represents the standard deviation of the data.

The Q_{ox}(Figure 3-1(a)) for the different splits reveals that there is oxide charge introduced during the post CMP anneal, regardless of whether the wafer was exposed to any contaminants or not. However, the wafers that were exposed to the dielectric CMP had more fixed oxide charge than the rest, indicating some level of contamination, while the wafers exposed to metal CMP match the annealed controls.

The D_{it} measurements indicate that the anneal also causes more interface traps. The wafers that underwent metal CMP appear to have 35% more interface traps than the wafers that underwent dielectric CMP, but the anneal alone caused a 129% increase, so this may be considered negligible.

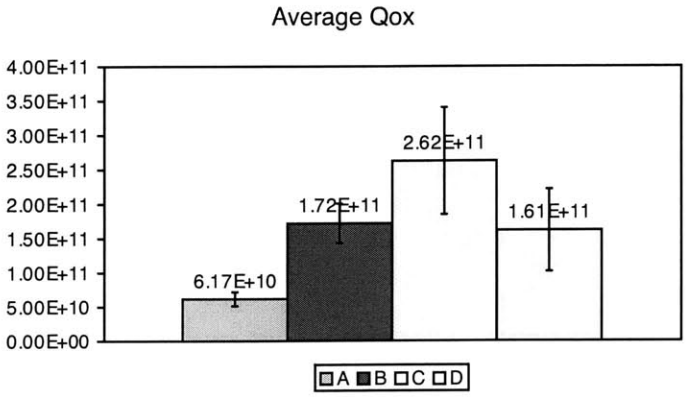
The T_s measurements show that the carrier lifetimes are all within a standard deviation of one another, except for the control wafers that underwent an anneal. This can be explained by the statistical variation of carrier lifetimes from wafer-to-wafer used for this experiment, which was higher than any impact made by processing.

The conclusion that can be made from this experiment is that while there may be more contaminant-based impact from using metal CMP compared to dielectric CMP, the impact is still on the same order.

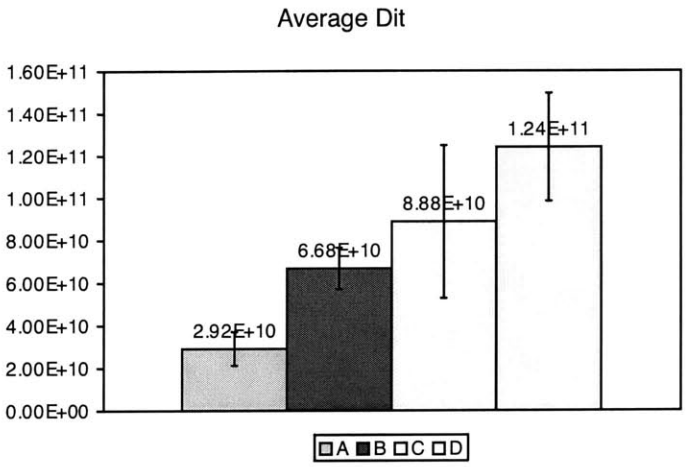
3.2.3 Cross-Contamination Effects

While it was found that there may be some risk to devices exposed to metal CMP that are subsequently processed at high temperatures, it was important to ascertain whether there is any risk of cross-contaminating other wafers that are processed in the same equipment. It was determined that the greatest risk of cross-contamination is in diffusion tubes, where the high temperatures make contaminant ions more mobile. With that in mind, a cross-contamination study was undertaken.

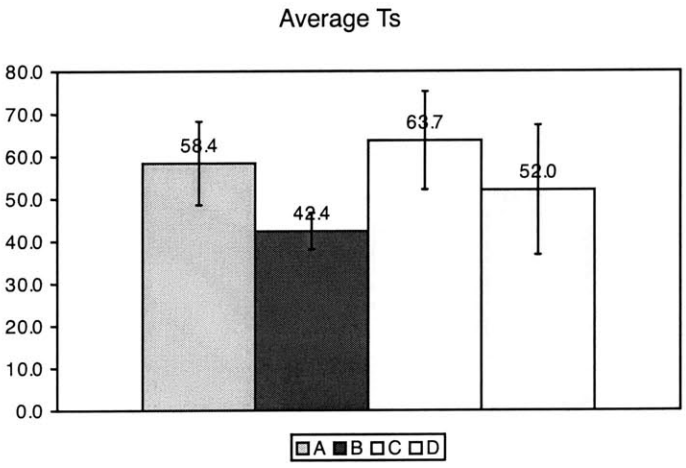
In this experiment, wafers were measured using the SCA, annealed at 850°C face-



(a) Fixed oxide charge



(b) Interface traps



(c) Minority carrier lifetimes

Figure 3-1: Contamination Effects Experimental Data

to-face with wafers that had been exposed to metal CMP, then measured again. The results are summarized in Figure 3-2. The bars represent the average of five measurements done on each wafers, where each group consists of the data for one wafer, before and after anneal.

In all cases, the results indicate that there is no difference between the wafers before and after annealing it next to a "contaminated" wafer. The measurements are all well within the standard deviation of the data, and are within the repeatability of the instrument.

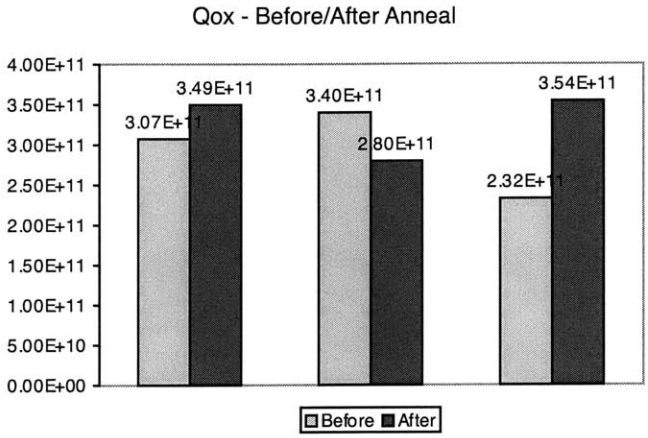
The conclusion that can be drawn from this experiment is that while there may be some risk to a wafer that is exposed to metal CMP, there is little or no risk associated to using a diffusion tube that has been used to process a wafer that has been exposed to metal CMP.

3.3 Implementation

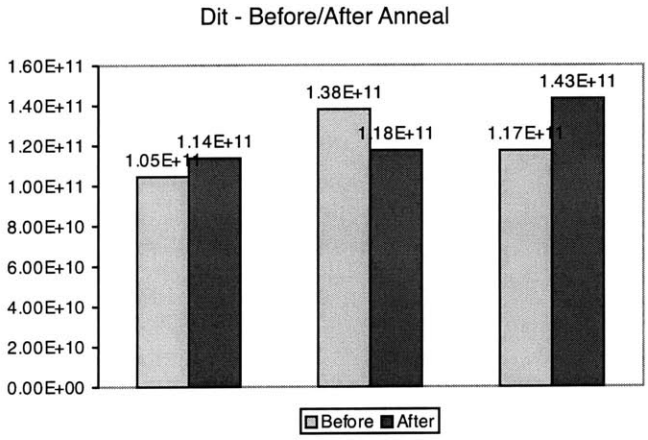
3.3.1 Process Flow

A simple DSI process was developed as a proof-of-concept experiment (Figure 3-3) to determine whether there are any processing difficulties inherent to the process or if device damage occurs during the process.

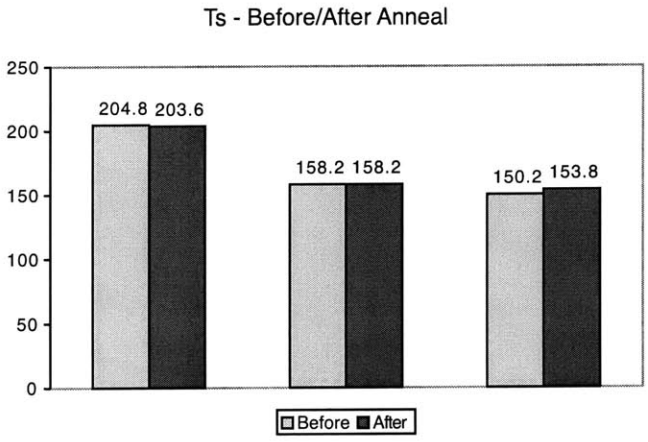
First, MOSFETs and test structures are constructed on an SOI wafer. Then, dual damascene tungsten interconnects are added. A dual-damascene technique is used for planarity, while tungsten is used as an interconnect material due to its resistance to subsequent high temperature steps. Then, a bonding layer of LTO is added (Figure 3-3(a)). The SOI wafer is then bonded to a handle wafer. The original SOI wafer is then removed, leaving only the SOI film and buried oxide (Figure 3-3(b)). PECVD oxide is deposited as an interlayer dielectric (Figure 3-3(c)), then a final layer of subtractive aluminum interconnects is added (Figure 3-3(d)).



(a) Fixed oxide charge

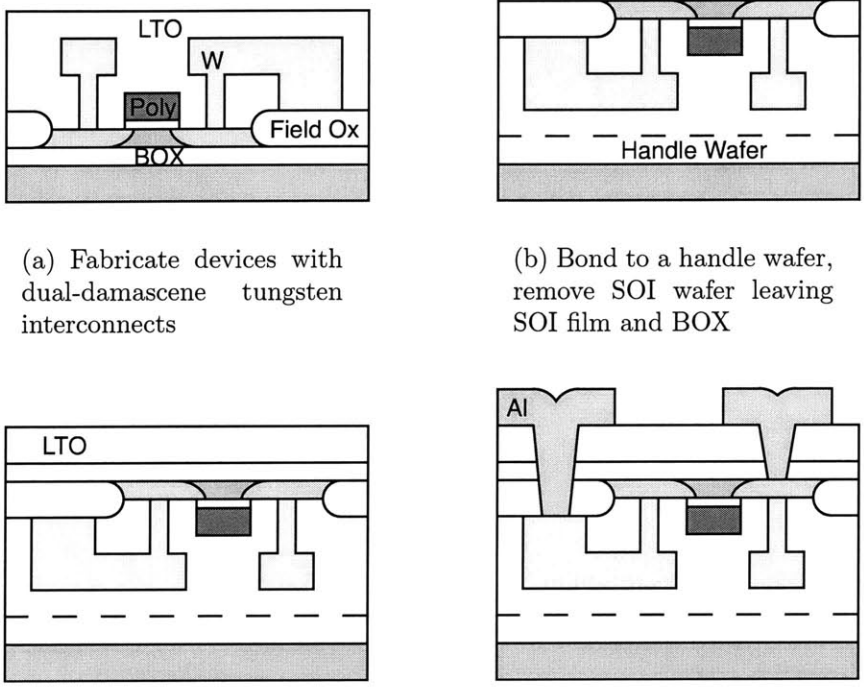


(b) Interface traps



(c) Minority carrier lifetimes

Figure 3-2: Cross-Contamination Experimental Data



(a) Fabricate devices with dual-damascene tungsten interconnects

(b) Bond to a handle wafer, remove SOI wafer leaving SOI film and BOX

(c) Deposit interlayer dielectric

(d) Build aluminum interconnects

Figure 3-3: Double-Sided Interconnect Process

3.3.2 Test Devices

Fully-depleted SOI NMOS test devices were constructed using the above procedure in order to determine whether device damage occurs during the DSI process. FETs built with all contacts made from the top level of metal functioned normally, as did devices made with the gate contacted from the bottom level of metal. However, the devices built with all contacts made from the bottom level of metal did not function properly. The source/drain contacts were open, meaning no current would flow through them. This is likely a result of the high-temperature processing after the contacts were made. During the bonding process, the wafers were heated to 850°C for six hours. During this time, it is likely that all of the silicon near the contacts diffused into the contact metal leaving a void between the contacts and the source/drain region. The reason that the poly-silicon contacts remained intact is that the poly-silicon thickness (300nm) is much greater than the silicon film thickness at the source/drain (50nm), allowing for more silicon loss before an open is formed. This problem indicates that either a method must be found to preserve the metal/silicon interface at elevated temperatures, or a lower thermal budget must be used.

3.4 Conclusions

DSI has the potential to increase interconnect performance and enable future 3D integration research. There are concerns as to whether the contamination risk of performing additional front-end processing on wafers that have already undergone back-end processing outweighs the advantages that DSI offers. The devices built successfully using this method indicate that this risk may be minor. However, the devices made by contacting the silicon before bonding were unsuccessful, indicating that this method requires more research before any widespread implementation.

Chapter 4

Extension to 3D

In a conventional CMOS circuit, there is a single layer that contains MOSFETs. This is referred to as a two-dimensional (2D) approach. While 2D circuits are highly successful, it is useful to explore alternative geometries to attempt to increase performance. In a three-dimensional (3D) circuit, there would be more than one level of MOSFETs stacked on top of each other.

4.1 Motivation and Concerns

4.1.1 Motivation

3D integrated circuits have several possible advantages that make them attractive for further research. These include packing density, alternative routing paths, and heterogeneous integration.

Compared to the finished packaging of a chip, a MOSFET is essentially 2D. While even the thinnest packaging is a few hundred microns thick, the thickness of the MOSFET itself is only a few thousand angstrom, and the entire circuit including interconnects is still only a few microns thick. By stacking many thin layers of circuits on top of each other, the space wasted by the packaging process can be reclaimed, increasing the final volume density of devices. For applications where form-factor is important, such as portable electronics, increased volume density is

highly advantageous.

Another advantage of 3D integrated circuits is allowing alternate routing paths. By stacking relevant circuit blocks on top of each other, chip area can be decreased, and global and semi-global interconnects can travel a short vertical distance rather than a long horizontal one [29, 30, 31]. The result is reduced parasitic resistance and capacitance, yielding reduced power requirements and/or increased performance.

3D integrated circuits can also improve system performance through heterogeneous integration. Any layer of circuits could potentially be of a different type, fabricated in a process optimized for those particular devices. CMOS logic, CMOS memory, bipolar RF circuits, and III-V optoelectronics could be fabricated under optimal conditions, even in separate fabrication facilities, then all integrated into the same chip [32, 33]. 3D integration could enable previously impossible system-on-a-chip (SOC) applications. An application for heterogeneous 3D integrated circuits proposed by Koyanagi et al. is computer vision chips. By integrating a CCD with memory and image processing circuitry, the bandwidth bottlenecks associated with having these devices on separate chips can be eliminated [34].

4.1.2 Concerns

An important issue surrounding 3D integration is yield. In a wafer-to-wafer integration scheme, and defective die will ruin the functionality of the die it is bonded to, making the entire circuit defective. This defect propagation problem means that yield must be very high for each wafer in order to have a reasonable yield of final circuits, prompting concern for the wafer-to-wafer integration scheme. Alternative schemes using known-good chips in chip-to-chip and chip-to-wafer integration have been suggested [35].

Despite the concern for defect propogation, there are several reasons that wafer-to-wafer integration may achieve acceptable yield. One potential advantage of 3D circuits is reduced die size. Each defect on the wafer can potentially ruin the die that it is located in. By reducing the die size, each wafer can have more dies. For an equal number of defects per wafer, the wafer with reduced die size will have a higher yield.

Another advantage is that each layer can be simpler. Rather than building BiCMOS circuits with DRAM on the same 2D chip, each layer in a 3D circuit can be built using a different process. Yield would be higher for building bipolar, NMOS, PMOS, and DRAM in separate processes rather than together.

Also, the back-end on each layer can be simpler. Complicated multi-layer interconnects result in reduced yield for 2D circuits. Each layer in a 3D circuit would likely require fewer levels of interconnects. Each layer would likely not require clock and power distribution, and global interconnects could be shared between levels.

Another concern surrounding 3D integration is heat removal. Heat dissipation is a problem in high-performance 2D logic circuit applications, and would be exacerbated by stacking additional layers od circuits on top of each other. While this will certainly be a problem for high-performance logic, the reduced RC delays of the interconnects reduce the power requirements of the circuit, thereby helping to alleviate the heat dissipation issue. Other potential applications for 3D integrated circuits where heterogeneos integration is the chief advantage may not be affected by heat removal issues.

4.2 Strata

This work suggests a 3D integration approach based on interchangeable building blocks, referred to as strata, where each stratum contains a set of devices and interconnects. By integrating these strata together, a 3D integrated circuit can be fabricated in a modular fashion.

There are three possible configurations of these strata. The lowest stratum (stratum 0) would contain the supporting substrate and a set of devices and interconnects. The devices may be bulk, SOI, SOI with DSI, or III-V. All connection to the rest of the strata and the outside world takes place through this level's top level of interconnects. A cross sectional view of an example for strata 0 is given in Figure 4-1.

The final stratum (stratum N) could be silicon devices with DSI, optoelectronic devices, microelectromechanical systems (MEMS), or any other integrated system

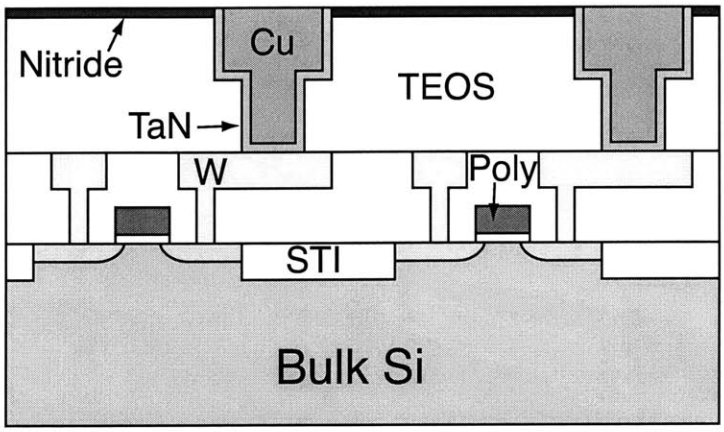


Figure 4-1: Example of Stratum 0

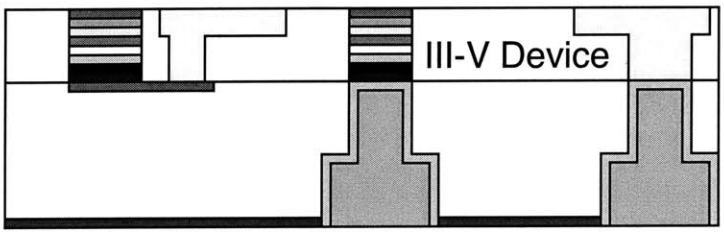


Figure 4-2: Example of Stratum N

suitable for wafer-bonding. Special measures may need to be taken to ensure the signals from lower levels can reach the outside world, but this would be implementation specific. Silicon devices with DSI have this capability built in. An example for stratum N with optoelectronic devices is given in Figure 4-2.

The intermediate strata (strata 1 to N-1) would be composed of SOI devices with DSI. These strata could contain digital logic, memory, and analog circuitry. For optimum performance, DG MOSFETs could be used. These intermediate strata would also pass signals from adjacent strata. An example of an intermediate stratum is given in Figure 4-3.

While most research is centered around integrating strata made from SOI (or even thinned bulk) devices with interconnects on a single side of the devices, strata built around DSI have the advantages of simpler integration and symmetry.

Integration schemes built around devices with interconnects on one side of the devices require additional back-end processing to make connections between the strata. As will be shown in Section 4.3, connecting strata with DSI only requires a bond and

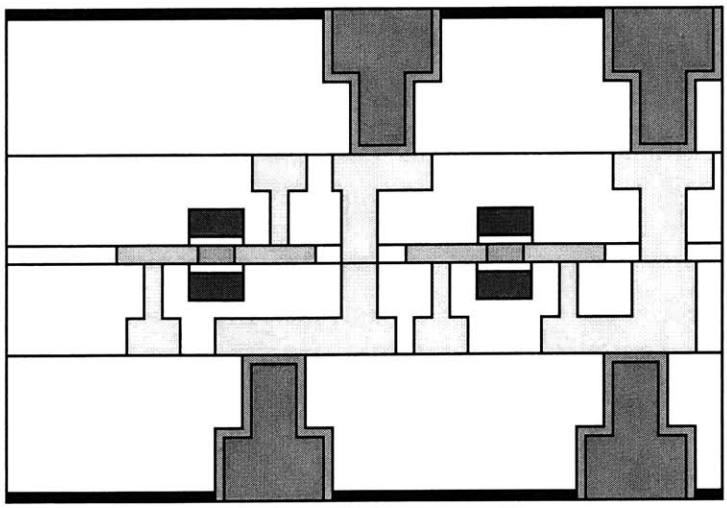


Figure 4-3: Example of an Intermediate Stratum

etch-back process.

Devices with DSI also have the advantage of offering symmetry around the device layer. Figure 4-4 shows a front-to-back 3D integration scheme. It is apparent that signals from lower strata must travel up and around to reach the devices on the next stratum. In this example, the first (local) level of interconnects on the second strata must be connected to the last (global) level of the first stratum. While minimizing the wasted vertical travel, the interconnection between global and local is most likely not desirable, and requires any additional levels of interconnects to be built after the bond. If the inter-strata vias are made between the global level on both strata, the signal must travel much further up and around. In a scheme with DSI, the signals can travel directly between global levels in straight lines.

4.3 3D Integration

Once the strata are constructed, integrating them into a 3D circuit is just a simple process of repeated bond and etch-back. Using aligned copper-to-copper bonding [36], the mechanical and electrical integration of the devices occurs simultaneously. Stratum 1 is bonded to stratum 0 (Figure 4-5), then the substrate of stratum 1 is etched away. More intermediate levels can be added, until the final stratum is added (Fig-

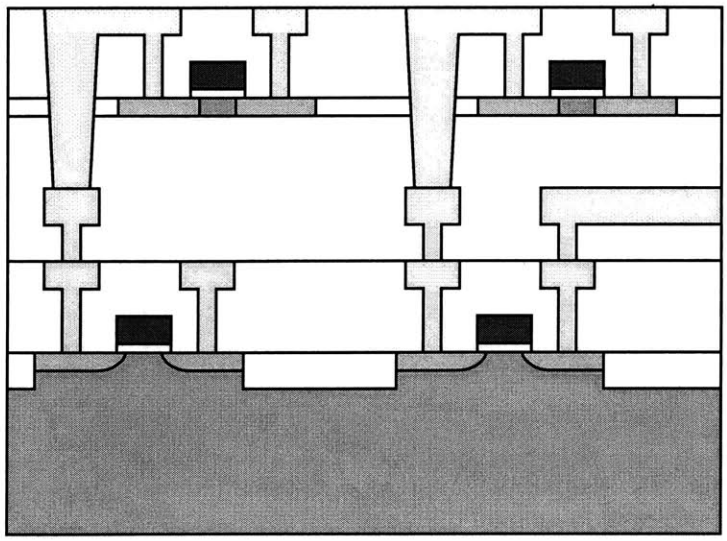


Figure 4-4: 3D Circuit Without DSI

ure 4-6). After removing the substrate from stratum N or perhaps some further processing, the 3D structure is complete (Figure 4-7).

In this example, the wafers are bonded together at a silicon nitride to silicon nitride interface rather than a silicon dioxide to silicon dioxide interface. Since silicon nitride is an effective diffusion barrier for copper, this means that misalignment during the bond will not leave and copper to silicon dioxide interfaces that would allow copper to diffuse into the device layer. Silicon nitride bonding is similar to silicon dioxide bonding, and good bond strength can be achieved at temperatures as low as 90°C [37].

4.4 Conclusions

3D integration has a great deal of potential for increasing the performance and versatility of integrated circuits. The technique presented here offers potential advantages over other proposed techniques. By fabricating 3D circuits using the building blocks of double-gate devices and DSI, higher-performance circuits with higher packing density can be achieved than with current techniques.

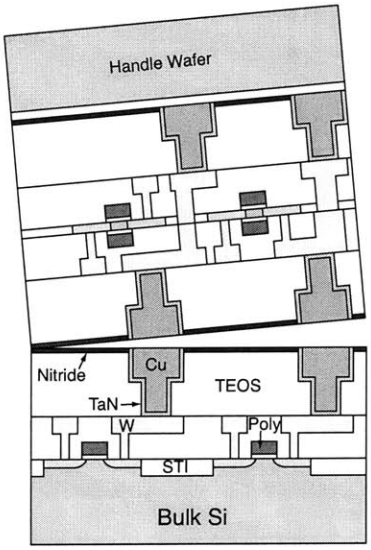


Figure 4-5: Bond Strata 1 to Strata 0

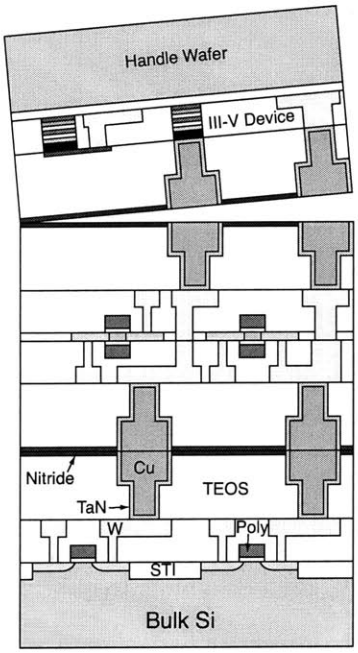


Figure 4-6: Bond Strata N to Strata N-1

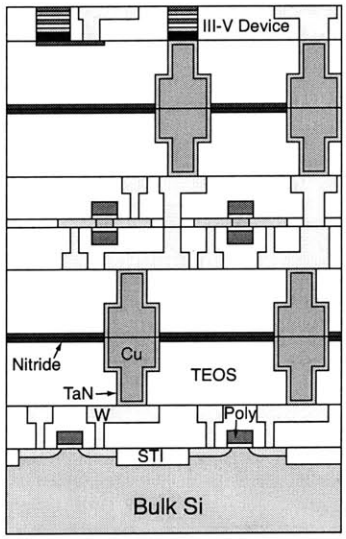


Figure 4-7: Completed 3D Circuit

Bibliography

[1] Yuan Taur et al. CMOS scaling into the nanometer regime. *Proceedings of the IEEE*, 85(4):486, April 1997.

[2] D.J. Frank et al. Monte Carlo simulation of a 30 nm dual-gate MOSFET: How short can Si go? *IEDM*, page 553, 1992.

[3] Andy Wei and Dimitri Antoniadis. Minimizing floating-body-induced threshold voltage in partially depleted SOI CMOS. *IEEE Electron Device Letters*, 17(8):391, August 1996.

[4] C. Wann et al. High-performance 0.01-um CMOS with a 9.5-ps gate delay and 150 ghz ft. *IEEE Electron Device Letters*, 18(12):625, December 1997.

[5] Hon-Sum Philip Wong et al. Nanoscale CMOS. *Proceedings of the IEEE*, 87(4), April 1999.

[6] Hon-Sum Philip Wong et al. Self-aligned (top and bottom) double-gate MOSFET with a 25 nm thick silicon channel. *IEDM*, page 747, 1997.

[7] Hon-Sum Philip Wong et al. Self-aligned (top and bottom) double-gate MOSFET with a 25nm thick silicon channel. *IEDM*, page 427, 1997.

[8] J.M. Hergenrother et al. The Vertical Replacement-Gate (VRG) MOSFET: A 50-nm vertical MOSFET with lithography-independent gate length. *IEDM*, page 75, 1999.

[9] Sang-Hyun Oh et al. 50 nm Vertical Replacement-Gate (VRG) pMOSFETs. *IEDM*, page 65, 2000.

[10] Digh Hisamoto et al. Impact of the vertical SOI "DELTA" structure on planar device technology. *IEEE Transactions on Electron Devices*, 38(6):1399, July 1991.

[11] Digh Hisamoto et al. A folded-channel MOSFET for deep-sub-tenth micron era. *IEDM*, page 1032, 1998.

[12] Xuejue Huang et al. Sub 50-nm FinFET: PMOS. *IEDM*, page 67, 1999.

[13] Digh Hisamoto et al. FinFET - A self-aligned double-gate MOSFET scalable to 20 nm. *IEEE Transactions on Electon Devices*, 47(12):2320, December 2000.

[14] C.-K. Hu and J.M.E. Harper. Copper interconnections and reliability. *Elsevier Materials Chemistry and Physics*, 52:5, 1998.

[15] Shyam P. Murarka and Steven W. Hymes. Copper metallization for ULSI and beyond. *Critical Reviews in Solid State and Materials Sciences*, 20(2):87, 1995.

[16] A.G. Milnes. *Deep Impurities in Semiconductors.* John Wiley and Sons, New York, 1973.

[17] Bert Vermeire et al. The effect of copper contamination on filed overlap edges and perimeter junction leakage current. *IEEE Transactions on Semiconductor Manufacturing*, 11(2):232, 1998.

[18] J. Torres et al. Study of Cu contamination during copper integration is a dual-damascene architecture for sub-quarted micron technology. *IITC*, page 232, 1998.

[19] J. D. McBrayer et al. Diffusion of metals in silicon dioxide. *Journal of the Electrochemical Society*, 133(6):1242, 1986.

[20] Ki-Bum Kim. The diffusion barrier property of TiN and TaN for Cu metallization. *Conference Proceedings ULSI XI*, page 281, 1996.

[21] S. Lakshminarayanan et al. Contact and via structures with copper interconnects fabricated using dual damascene technology. *IEEE Electron Device Letters*, 15(8):307, August 1994.

[22] Hon-Sum Wong et al. Design and performance considerations for sub-0.1um double-gate SOI MOSFET's. *IEDM*, page 747, 1994.

[23] Tetsu Tanaka et al. Analysis of P+ poly Si double-gate thin-film SOI MOSFETS. *IEDM*, page 683, 1991.

[24] Andrew Ritenour. Fabrication technology for double-gate field effect transistors. Master's thesis, Massachusetts Institute of Technology, 1999.

[25] K. Terada and H. Muta. A new method to determine effective mosfet channel length. *Japanese Journal of Applied Physics*, 18:935, 1979.

[26] Robert F. Pierret. *Semiconductor Device Fundamentals.* Addison-Wesley Publishing Company, Reading, MA, 1996.

[27] Avid Kamgar et al. Reduced electron mobility due to nitrogen implant prior to the gate oxide growth. *IEEE Electron Device Letters*, 21(5):227, 2000.

[28] Yuan Taur and Tak H. Ning. *Fundamentals of Modern VLSI Design.* Cambridge University Press, New York, 1998.

[29] Arifur Rahman et al. Wire-length distribution of three-dimensional integrated circuits. *IITC*, page 233, 1999.

[30] Arifur Rahman et al. Comparison of key performance metrics in two- and three-dimensional integrated circuits. *Proceedings of the IEEE*, page 18, 2000.

[31] Jeffrey Davis et al. Interconnect limits on gigascale integration (GSI) i the 21st century. *Proceedings of the IEEE*, 89(3):305, 2001.

[32] Joanna London et al. Preparation of silicon-on-gallium arsenide wafers for monolithic optoelectronic integration. *IEEE Photonics Technology Letters*, 11(8):958, 1999.

[33] Mitsuo Matsunami et al. *International Conference on Wafer Scale Integration*, page 322, 1990.

[34] Mitsumasa Koyanagi et al. Neuromorphic vision chip fabricated using three-dimensional integration technology. *ISSCC*, page 270, 2001.

[35] Dirk J. Bartelink. Integrated systems. *IEEE Transactions on Electron Devices*, 43(10):1678, 1996.

[36] Andy Fan et al. Copper wafer bonding. *Electrochemical and Solid-State Letters*, 2(10):534, 1999.

[37] Rebert Bower et al. Low temperature Si_3N_4 direct bonding. *Applied Physics Letters*, 62(26):3485, 1993.