Silicon VLSI Technology

Fundamentals, Practice and Modeling

James D. Plummer
Michael Deal
Peter B. Griffin
Department of Electrical Engineering
Stanford University

Prentice Hall
Upper Saddle River, NJ 07458

Kurt K. Wendt Library
University of Wisconsin-Madison
215 N. Randall Avenue
Madison, WI 53706-1688

Library of Congress Cataloging-in-Publication Data
Silicon VLSI technology
p. cm.
ISBN 0-13-085037-3
1. Integrated circuits—Very large scale integration—Design and construction. 2. Silicon. 3. Silicon oxide films. 4. Metal oxide semiconductors. 5. Silicon technology.
TK7874.75.S54 2000
621.39'5—dc21 99-42745
CIP

Publisher: Tom Robbins
Associate Editor: Alice Dworkin
Editorial/Production Supervision: Rose Kernan
Vice President and Editorial Director, ECS: Marcia Horton
Vice President of Production and Manufacturing: David W. Riccardi
Executive Managing Editor: Vince O'Brien
Marketing Manager: Danny Hoyt
Managing Editor: David A. George
Manufacturing Buyer: Pat Brown
Manufacturing Manager: Trudy Pisciotti
Art Director: Jayne Conte
Cover Design: Bruce Kenselaar
Editorial Assistant: Jesse Power
Copy Editor: Martha Williams
Composition: D&G Limited, LLC

©2000 by Prentice Hall, Inc.
Upper Saddle River, New Jersey 07458

All rights reserved. No part of this book may be reproduced, in any form or by any means, without permission in writing from the publisher.

The author and publisher of this book have used their best efforts in preparing this book. These efforts include the development, research, and testing of the theories and programs to determine their effectiveness. The author and publisher make no warranty of any kind, expressed or implied, with regard to these programs or the documentation contained in this book. The author and publisher shall not be liable in any event for incidental or consequential damages in connection with, or arising out of, the furnishing, performance, or use of these programs.

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-085037-3

Prentice Hall International (UK) Limited, *London*
Prentice Hall of Australia Pty. Limited, *Sydney*
Prentice Hall Canada Inc., *Toronto*
Prentice Hall Hispanoamericana, S.A., *Mexico*
Prentice Hall of India Private Limited, *New Delhi*
Prentice Hall of Japan, Inc., *Tokyo*
Pearson Education Pte., Ltd. , *Singapore*
Editora Prentice† Hall do Brasil, Ltda., *Rio de Janeiro*

Back-End Technology

11

11.1 Introduction

Back-end technology refers to the interconnect layers, contacts, vias and dielectric layers that wire the active devices into specific circuit configurations. Figure 11–1 is a schematic diagram taken from Chapter 2 showing these components in a typical integrated circuit structure. Figure 11–2 is a SEM micrograph of an IBM circuit showing the back-end structure with the dielectric layers etched away, revealing the interconnects, contacts, and vias. Another SEM of this same chip was shown in Figure 1–12. The relative importance of back-end structures has greatly increased in recent years because the circuit delays associated with interconnects have not kept pace with the faster device speeds provided by scaled technologies. Thus much recent effort has been devoted to improving back-end technology and it is widely believed that these efforts will only accelerate in the future.

Interconnects can be either local or global. In general, local interconnects are the first, or lowest, level of interconnects. They usually connect gates, sources, and drains in MOS technology, and emitters, bases, and collectors in bipolar technology. In MOS technology a local interconnect, polycrystalline silicon, also serves as the gate electrode material. Silicided gates and silicided source/drain regions and materials like TiN and W can also act as local interconnects. Local interconnects can tolerate higher resistivities than global interconnects since they are not very long. But they must also, in most cases, be able to withstand higher processing temperatures because they are deposited earlier in the process flow than global interconnects. Global interconnects, mostly made of Al, are generally all the interconnect levels above the local interconnect level. They often cover large distances between different devices and different parts of the chip and therefore are always low-resistance metals.

Ohmic contacts connect an interconnect with active regions or devices in the silicon substrate. A high-resistivity dielectric layer, usually silicon dioxide, separates the active regions from the first-level global interconnect, and electrical contact is made between the interconnect and the active regions in the silicon through openings in that dielectric layer. At the same time in the processing, contact can be made between the first-level global interconnect and the local interconnects, since they are also separated by that same dielectric layer. Connections between two levels of global interconnects are usually given a different name—vias. They are made through openings in the different levels of the intermetal dielectrics (IMDs), which separate interconnects from each other and from the active areas and devices. The bottom dielectric layer through which the contacts are made is usually called the first-level dielectric. Examples of these structures are shown in Figures 11–1 and 11–2.

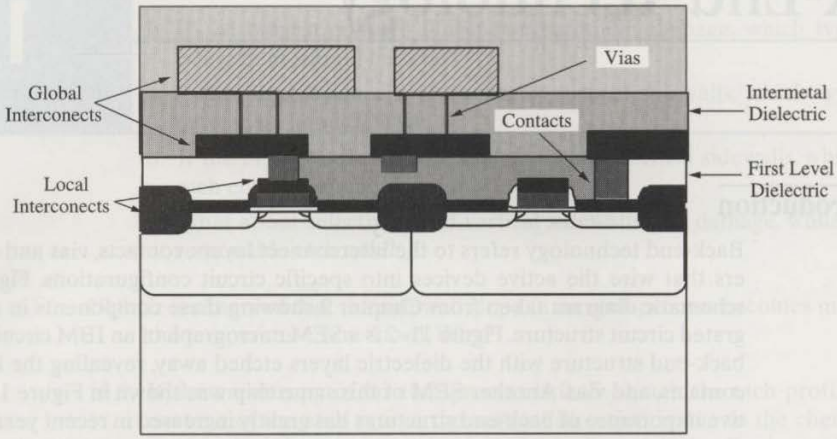

Figure 11–1 Schematic cross section of back-end structure, showing interconnects, contacts, and vias, separated by dielectric layers (cross-hatched regions).

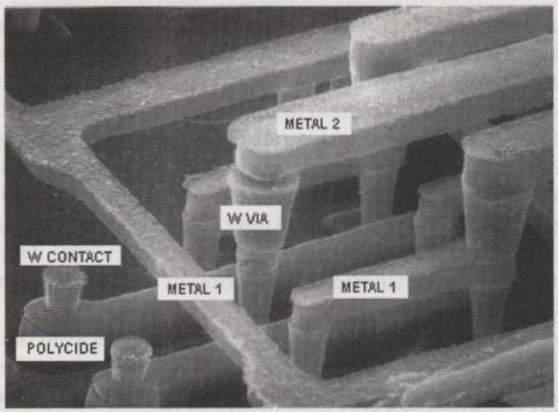

Figure 11–2 SEM micrograph of an IBM circuit showing the interconnects with the dielectric layers etched away. Photograph courtesy of Integrated Circuit Engineering Corporation.

It is easy to see why interconnect structures are becoming more important in overall chip performance. Consider the structure in Figure 11–3 [11.1]. The *RC* time delay of a signal propagating along such a line is to first order obtained by treating it as a distributed, unterminated transmission line. For such a system the time delay, τ_L, is approximately equal to $0.89RC$ where R is the line resistance and C is the total capacitance associated with the line [11.2]. A more accurate analysis would include such elements as the load capacitance, driver resistance, and line inductance [11.3, 11.4, 11.5]. However, the $0.89RC$ expression for the time delay which considers only the interconnect *RC* component is sufficient for our purposes.

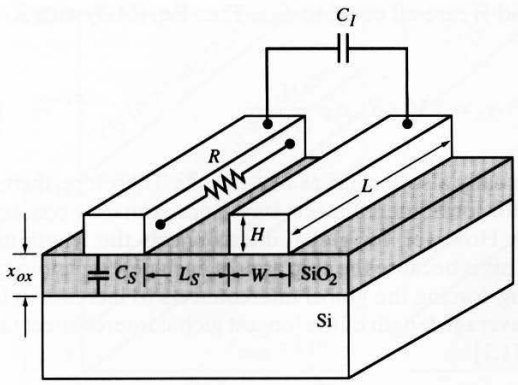

Figure 11-3 Interconnect structure for *RC* analysis. The two lines on top are the metal interconnects of dimensions *W*, *L*, and *H*, sitting on a SiO_2 layer. There is also SiO_2 between and above the metal lines. (After [11.1].)

The line resistance, in Ω, of one of the interconnects is given by

$$R = \rho \frac{L}{WH} \tag{11.1}$$

where ρ is the interconnect resistivity, and *L*, *W*, and *H* are the interconnect length, width, and height, respectively. The total capacitance associated with the line is

$$C = K_{ox}\varepsilon_o \frac{WL}{x_{ox}} + K_{ox}\varepsilon_o \frac{HL}{L_S} \tag{11.2}$$

where x_{ox} and K_{ox} are the oxide thickness and dielectric constant, respectively, and ε_o is the permittivity of free space. The first term represents the line to substrate capacitance, C_S, and the second term the coupling capacitance between adjacent lines, C_I. It is assumed that the lines are surrounded by oxide on all sides.

The total *RC* delay associated with the line is thus

$$\tau_L = 0.89 \cdot K_I K_{ox}\varepsilon_o \rho L^2 \left(\frac{1}{Hx_{ox}} + \frac{1}{WL_S}\right) \tag{11.3}$$

where K_I is added to empirically account for fringing fields and other interconnects above and below the line in multilayer interconnect systems. K_I is often taken to be ≈ 2.

Now we must consider what happens to the dimensions of the structure as the technology progresses. First, L_S and *W* are often close to the minimum feature size F_{min}. As the technology progresses, F_{min} gets smaller. The values of L_S and *W* may in fact be larger than the minimum feature size for some interconnect levels, but they still commonly scale with F_{min}. Both x_{ox} and H also shrink as F_{min} shrinks in this ideal scaling scheme, keeping the aspect ratio, *H/W*, constant. In the original analysis by Saraswat and Mohammadi [11.1], it was assumed that x_{ox} and *H* equaled $0.35F_{min}$ and $0.25F_{min}$, respectively. However, these thicknesses have not decreased as quickly as this in recent years, especially for the global interconnects. To keep the analysis simple, we assume

that x_{ox} and H, as well as L_S and W, are all equal to F_{min}. Thus Eq. (11.3) with K_I equal to 2, becomes

$$\tau_L = 3.56 \cdot K_{ox}\varepsilon_o\rho \frac{L^2}{(F_{min})^2} \tag{11.4}$$

For local interconnects, L usually also shrinks as F_{min} shrinks. Therefore, the net result is that the RC time delay for local interconnects stays approximately constant according to this scaling scheme. However, for global interconnects the length usually increases rather than shrinks. This is because the chip area of each new technology generation usually keeps increasing, forcing the global interconnects to increase in length to connect all the devices. The average length of the longest global interconnect in a circuit can be approximated by [11.1]

$$L_{max} = \frac{\sqrt{A}}{2} \tag{11.5}$$

where A is the chip area. Inserting this into Eq. (11.4) finally leads to the following expression for the delay associated with global interconnects:

$$\tau_L = 0.89 \cdot K_{ox}\varepsilon_o\rho \frac{A}{(F_{min})^2} \tag{11.6}$$

Example Calculate the interconnect delay time according to Eq. (11.6) for the following conditions: Al interconnect ($\rho = 3.0 \times 10^{-6}\ \Omega$ cm), SiO_2 dielectric ($K_{ox} = 3.9$), area $= 100$ mm^2, and the minimum feature size of 0.35 μm.

Answer

$$\tau_L = 0.89 \cdot K_{ox}\varepsilon_o\rho \frac{A}{(F_{min})^2}$$

$$= 0.89 \cdot 3.9 \cdot 8.86 \times 10^{-14}\ \text{Fcm}^{-1} \cdot 3.0 \times 10^{-6}\ \Omega\ \text{cm} \frac{1\text{cm}^2}{(0.35 \times 10^{-4}\ \text{cm})^2}$$

$$= 7.5 \times 10^{-10}\ \text{sec} = 0.75\ \text{nsec}$$

As technology progresses, F_{min} shrinks and A usually gets larger. Both of these lead to an increase in the RC time delay in global interconnects. Even if A stays constant, the time delay will increase due to F_{min} shrinking. Physically this occurs because as the interconnect cross-sectional area decreases, R increases, but C does not decrease in proportion because the dielectric thicknesses surrounding the line also decrease. Hence RC

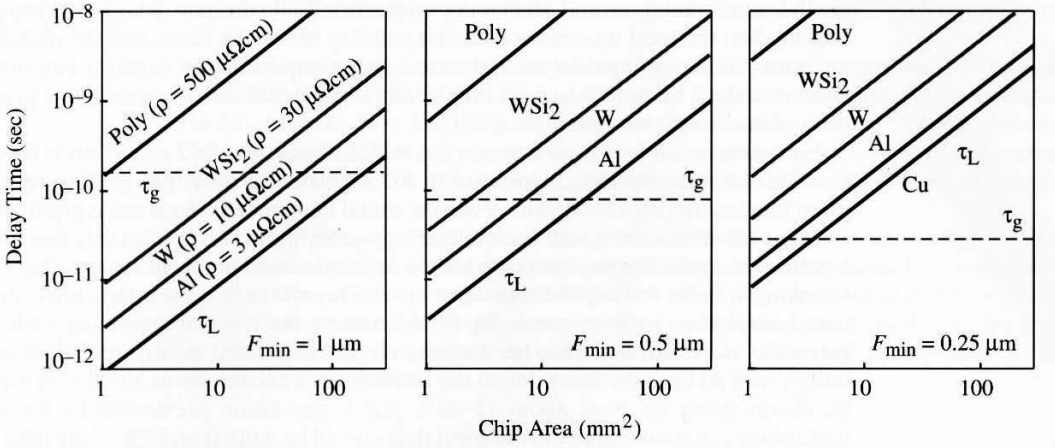

Figure 11–4 Global interconnect and gate time delay versus chip area and minimum feature sizes, F_{min}, for various interconnect materials. Interconnect delay corresponds to longest global interconnect in the circuit. (Updated from [11.1].)

increases. Having A increase only exacerbates this problem. Figure 11–4 plots Eq. (11.6) as a function of chip area for different values of F_{min} and different interconnect materials, updated from the original figure in [11.1] to use the new expressions for x_{ox} and H and to include the F_{min} = 0.25 μm case. Cu is also included as a possible interconnect material in the 0.25-μm technology. The interconnect delay here is based on the longest global interconnect using Eq. (11.6), with minimum feature size dimensions and assuming that the driver or load elements do not contribute significantly to the overall time delay. Also plotted for comparison is a typical gate delay, τ_g, for each minimum feature size, as measured in ring oscillator circuits. The gate delay is dependent only upon F_{min} and not A and thus decreases as the minimum feature size decreases. It is obvious that for larger F_{min}'s, the gate delay is generally large relative to the delay for Al global interconnects. (And one can also see why one would not want to use polysilicon as a global interconnect.) But for smaller minimum features and larger chip areas, the global interconnect delay becomes quite large compared to the gate delay and can have a large impact on the circuit performance.

In 0.25 μm (=F_{min}) manufacturing technology, chip areas are typically 200–350 mm^2 (corresponding to a maximum global interconnect length of 7–9 mm), and gate delays as measured in ring oscillators are generally about 30–70 picoseconds. The RC delay of the longest global interconnect line, with minimum feature size dimensions, by this analysis is about 3–5 nanoseconds for Al lines, or about 80 times the gate delay. This does not necessarily mean that circuit performance is totally dominated by the interconnects in today's integrated circuits. The specific circuit architecture, the exact heights, widths and lengths of the interconnect lines on the different levels, the driver and load components, and other circuit factors will determine the speed of the circuit and the relative importance of the front-end and back-end elements. What is important

are the trends that these and other analyses predict: While the gate delays will keep getting smaller, the local interconnect delays will stay about the same, and the global interconnect delays will get larger and have a larger impact on the circuit performance. Recent analysis by Bohr based on Intel's technology agree with the qualitative predictions of this simple analysis [11.6, 11.7].

Each new technology generation in the NTRS represents a $\sqrt{2}$ reduction in feature size. This corresponds to a 2x increase in *RC* interconnect delay per generation for a given line length [Eq. (11.4)]. However, the metal line aspect ratio is not typically kept constant but is increased with each technology generation. This is primarily due to the fact that the metal line thickness has stayed rather constant in recent times rather than shrinking in order to keep the resistance lower. This means that the interconnect delay does not increase quite as fast as Eq. (11.4) predicts, but it is still increasing while the gate delay is decreasing. It has been estimated [11.6, 11.8] that in current CMOS technology with Al lines the delay due to the interconnects can be about 30–40% of the total circuit delay, up from about 15–20% just a generation previously. In the next technology generations, the interconnect delay could be more than 50% of the total circuit delay, emphasizing the importance of the back-end structures.

As feature sizes shrink, the size of the contacts is also getting smaller. Since to first order the contact resistance, R_c, is inversely proportional to the contact area, the importance of the contacts increases as the circuit dimensions are shrunk.

The specific projections for back-end technology given by the NTRS are shown in Table 11–1. A number of the qualitative observations made earlier are apparent in this table. Note the increasing numbers of wiring levels in the future and the increasing aspect ratios associated with interconnect lines as well as contacts and vias. Finally, notice the decreasing metal resistivity requirement which cannot be met by aluminum in the future, and the decreasing dielectric constant required for intermetal dielectrics, which cannot be met by SiO_2. Clearly new materials will have to be introduced into back-end

Table 11–1 Future projections for back-end technology taken from the SIA NTRS [11.9]

Year of First DRAM Shipment	1997	1999	2003	2006	2009	2012
Minimum Feature Size, F_{min} (nm)	250	180	130	100	70	50
DRAM Bits/Chip	256M	1G	4G	16G	64G	256G
DRAM Chip Size (mm^2)	280	400	560	790	1120	1580
MPU Chip Size (mm^2)	300	360	430	520	620	750
Wiring Levels—Logic	6	6–7	7	7–8	8–9	9
Min Metal CD (nm)	250	180	130	100	70	50
Min Contact/Via CD nm	280/360	200/260	140/180	110/140	80/100	60/70
Metal Aspect Ratio	1.8	1.8	2.1	2.4	2.7	3.0
Contact Aspect Ratio (DRAM)	5.5	6.3	7.5	9	10.5	12
Via Aspect Ratio (logic)	2.2	2.2	2.5	2.7	2.9	3.2
Metal Resistivity ($\mu\Omega$-cm)	3.3	2.2	2.2	2.2	<1.8	<1.8
Interlevel Metal Dielectric Constant	3.0–4.1	2.5–3.0	1.5–2.0	1.5–2.0	<1.5	<1.5

structures to meet these objectives. Cu interconnects and "low K" dielectrics are the likely answer to these requirements.

Metallization structures are also receiving increased attention because of failure and reliability issues. Failures are due to a variety of causes: stress-induced void formation in interconnect lines; hillock growth; unwanted chemical reactions and interdiffusion between layers; dopant diffusion and segregation; excessive roughening of films; corrosion; electromigration; and others. Many of these will be discussed in more detail later in the chapter.

Finally it should be pointed out that in the schematic figures in this section we illustrated each metal or dielectric level as simple single-layer structures. In practice, each level or layer is often a multilayered structure itself. The trends to multilayered back-end structures and the use of new processing techniques have been due to the ever-increasing demands on both processing stability and electrical performance as the technology progresses.

11.2 Historical Development and Basic Concepts

The earliest integrated circuits in the 1960s consisted of aluminum lines connecting active regions—transistors and resistors—in silicon. Silicon dioxide passivated the silicon surface and isolated the aluminum interconnects from those regions in the substrate where electrical contact was not wanted. Where contact was desired, openings in the oxide were made so that the aluminum could come in contact with the silicon. Figure 11–5 shows an early structure.

Pure aluminum was thus used for both the contact material and the interconnect. Why aluminum? First, it has a low electrical resistivity at room temperature. (See Table 11–2 for a listing and properties of commonly used interconnect materials). Second, it adheres well to silicon and silicon dioxide. Third, it makes good electrical contact to heavily doped Si. The latter two characteristics are related to the fact that the oxide of Al is very stable, more stable than SiO_2. This means that if Al is deposited on top of SiO_2, it will react with it even at low temperatures, forming a thin layer of Al_2O_3 at the interface. This forms a glue layer, binding the Al to the SiO_2. In the case of Al-to-Si contacts, the Al will reduce any native Si oxide on top of the Si. Without this reaction, the ubiquitous native oxide on the silicon could prevent ohmic contact. Fortunately, the Al_2O_3 that is formed allows Al to diffuse through it into the Si, so that good ohmic contact between the Al and Si is obtained. A fourth benefit of using Al is that its presence

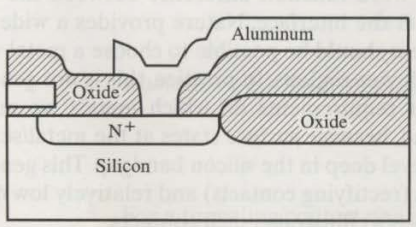

Figure 11–5 Early back-end structure with Al interconnect making contact to highly doped Si active region.

Table 11–2 Properties of interconnect materials

Material	Thin film resistivity ($\mu\Omega$cm)	Melting point (°C)
Al	2.7–3.0	660
W	8–15	3410
Cu	1.7–2.0	1084
Ti	40–70	1670
PtSi	28–35	1229
$TiSi_2$	13–16	1540
WSi_2	30–70	2165
$CoSi_2$	15–20	1326
NiSi	14–20	992
TiN	50–150	~2950
$Ti_{30}W_{70}$	75–200	~2200
polysilicon (heavily doped)	450–1000	1417

assists the annealing out of interface traps at the Si-SiO_2 interface, presumably by converting H_2O to free H, which passivates the traps. (See Chapter 6 where this process is described in more detail.) For all these reasons, contacts and interconnects made of aluminum were instrumental in the development of the first integrated circuits.

It was not long, though, before problems arose. As the circuits became more complicated, the structures became smaller, and the processing got more involved, new demands were placed on the metallization schemes. As a result, new metallization strategies, materials, and structures were developed. Even today the evolution continues, as evidenced by the changing requirements for interconnects in Table 11–1. We will now follow this evolution from the early days of integrated circuit fabrication, first with the contacts, then with the interconnects, and finally with the dielectrics.

11.2.1 Contacts

Contacts provide low-resistance connections between the metal interconnects and active device regions. Metal/semiconductor contacts can in general be rectifying (Schottky) or ohmic. Figure 11–6 shows the band diagrams for these modes of current transfer through a contact. Because of the work function difference between the metal and semiconductor, ϕ_B, a barrier exists at the interface. Nature provides a wide variety of metal work functions so in principle it should be possible to choose a metal which produces a small ϕ_B contact to silicon device regions. In practice, this is not generally possible, first because there are other requirements on which metals we may choose (resistivity, for example), and second, because surface states at the metal/semiconductor junction tend to pin the Fermi level deep in the silicon bandgap. This generally produces high ϕ_B contacts to N regions (rectifying contacts) and relatively low ϕ_B (ohmic) contacts to p-type silicon. Figure 11–6(a) illustrates such contacts.

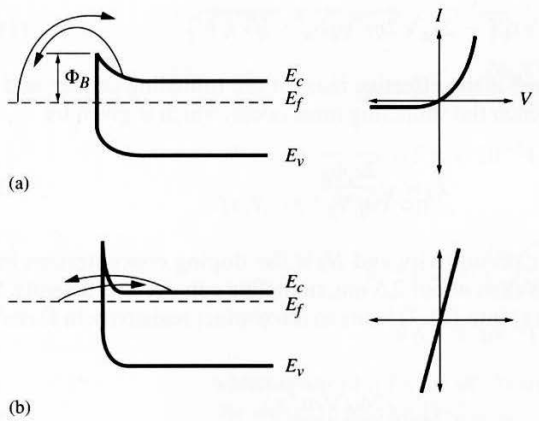

Figure 11–6 Band diagrams and I-V characteristics for metal/semiconductor junctions or contacts, illustrating different modes of current transfer: (a) Schottky or rectifying contact; (b) tunneling or ohmic contact.

Thermionic emission, shown in Figure 11–6(a), in which thermal energy allows the carriers to surmount the barrier, does allow some current to flow over a Schottky barrier. The basic properties of Schottky diodes operating in this mode are similar to the PN diode described in Chapter 1 in Section 1.4.1. However, for most metals the barrier is too high to n-type Si, and not enough current can flow under normal operating conditions.

The other way for the carriers to get past the barrier is to tunnel though it [Figure 11–6(b)]. If the silicon is heavily doped, the width of the barrier, defined by the depletion region, x_d, becomes narrow enough that many of the carriers can quantum-mechanically tunnel through the barrier. Ohmic behavior, in which current can flow virtually unimpeded in both directions through a contact, can thus be achieved. This occurs when the doping is greater than about 6×10^{19} cm^{-3}.

The specific contact resistivity, ρ_c, is defined as

$$\rho_c \equiv \left[\frac{\partial V_{MS}}{\partial J}\right]_{V_{MS}=0} \tag{11.7}$$

where J is the current density across the metal/semiconductor interface and V_{MS} is the voltage across it. For a Schottky contact governed by thermionic emission over the barrier, the current density is given by

$$J_S = A^* T^2 \exp\left(\frac{-q\phi_B}{kT}\right)\left(\exp\left(\frac{qV}{kT}\right) - 1\right) \tag{11.8}$$

where A* is Richardson's constant. The specific contact resistivity, in Ω cm^2, is therefore

$$\rho_c = \frac{k}{qA^*T}\exp\left(\frac{q\phi_B}{kT}\right) = \frac{kT}{qJ_s} \tag{11.9}$$

For a tunneling contact, the analysis is somewhat more complicated, resulting in

$$J_s \propto \exp[-2x_d\sqrt{2m^*(q\phi_B - qV)/\hbar^2}] \tag{11.10}$$

where $\hbar$ is Planck's constant, m^* is the effective mass of the tunneling carrier, and x_d is the depletion layer through which the tunneling must occur, which is given by

$$x_d = \sqrt{\frac{2\varepsilon_s\phi_B}{qN_d}} \tag{11.11}$$

where ε_s is the semiconductor permittivity and N_d is the doping concentration in the semiconductor. When x_d is less than about 2.5 nm, tunneling can occur efficiently. Substituting Eqs. (11.10) and (11.11) into (11.7) leads to the contact resistivity, in Ω cm^2, for tunneling contacts:

$$\rho_c = \rho_{co}\exp\left(\frac{2\phi_B\sqrt{m^*\varepsilon_s}}{\hbar\sqrt{N_d}}\right) \tag{11.12}$$

where ρ_{co} is a constant dependent on the metal and semiconductor. The contact resistivity goes down as $\sqrt{N_d}$ goes up because x_d shrinks and tunneling becomes more efficient. For x_d to equal 2.5 nm or less in Si with an Al contact and result in ohmic behavior, N_d must equal $\approx 6 \times 10^{19}$ cm^{-3}. The contact resistance, in Ω, is the contact resistivity divided by the contact area A, assuming that J is uniform over the contact area.

Figure 11–7 shows experimental data for specific contact resistivity as a function of N_d, along with the theoretical prediction based on the preceding equations. At low doping, thermionic emission controls the process which is essentially independent of doping, while at high doping, ρ_c is dominated by tunneling which is a very strong function of the doping level. Measured ρ_c values may be higher than indicated in Figure 11–7 if a clean interface is not obtained, due to the presence of unetched material or etch residues in the contact hole, native oxide present at the metal/semiconductor interface, or other contaminants or reaction products. Another common reason for the contact resistivity being higher than expected is that there are fewer dopants in the Si at the surface than desired. This may be due to segregation of dopants into the contact materials above the Si.

Example The expression for specific contact resistivity for high doping levels where tunneling dominates, Eq. (11.12), can be written in the form

$$\rho_c = \rho_{co}\exp\left(\frac{\phi_B C_1}{\sqrt{N_d}}\right)$$

where $C_1 \approx 7.0 \times 10^{10}$ cm$^{-3/2}$ eV^{-1} for Si. If $\phi_B = 0.6$ eV and $\rho_{CO} = 1.0 \times 10^{-7}\ \Omega \cdot$ cm^2 for a Si-Al contact, how much does the specific contact resistivity decrease when the doping is increased from 1×10^{19} to 1×10^{20} cm^{-3}?

Answer At $N_D = 1 \times 10^{19}\ \text{cm}^{-3}$,

$$\rho_c = \rho_{co} \exp\left(\frac{\phi_B C_1}{\sqrt{N_d}}\right) = 1.0 \times 10^{-7} \exp\left(\frac{0.6 \cdot 7.0 \times 10^{10}}{\sqrt{1 \times 10^{19}}}\right)$$

$$= 5.9 \times 10^{-2}\ \Omega\ \text{cm}^2$$

At $N_D = 1 \times 10^{20}\ \text{cm}^{-3}$,

$$\rho_c = \rho_{co} \exp\left(\frac{\phi_B C_1}{\sqrt{N_d}}\right) = 1.0 \times 10^{-7} \exp\left(\frac{0.6 \cdot 7.0 \times 10^{10}}{\sqrt{1 \times 10^{20}}}\right)$$

$$= 6.7 \times 10^{-6}\ \Omega\ \text{cm}^2$$

which is about a factor of about 9000 lower. These numbers are in agreement with the values in Figure 11–7.

The general requirements for metal-to-silicon contacts are low contact resistance and good thermal stability. Low contact resistance means a high-dopant concentration

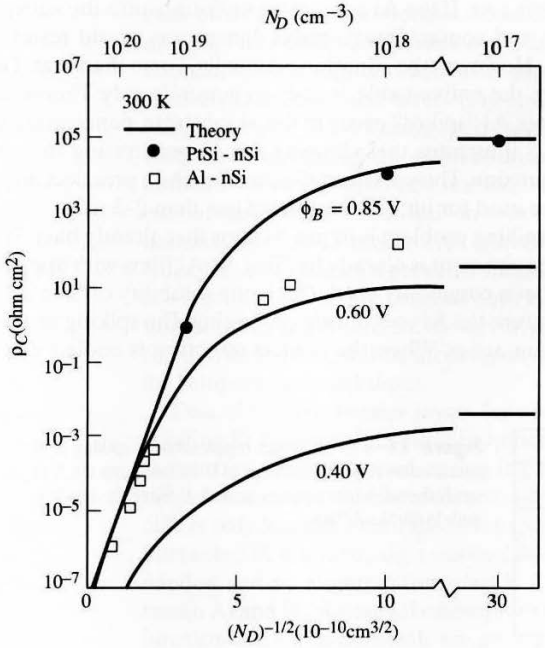

Figure 11–7 Contact resistivity (ρ_C) versus doping for three values of the metal/semiconductor work function. Both theoretical curves, based on Eqs. (11.9) and (11.12), and experimental data are plotted. (After [11.10, 11.11].)

at the Si surface and a good contact between the metal and silicon. Good contact requires an interface free of contaminants and residues. Good thermal stability means that the contact structure does not degrade during subsequent thermal processing, nor does it cause other areas to be adversely affected, such as active areas in the silicon underneath the contact.

A problem with thermal stability resulted in the first major change in the ohmic contact metallurgy. As mentioned previously, early integrated circuits used pure Al as the interconnect material, making direct contact to the Si substrate in heavily doped regions. Al makes a good contact to silicon by reducing the native oxide on the surface, which also removes any other contaminants on the surface. To ensure that the Al reduces the native Si oxide and that the Al is in good physical contact to the Si, a sintering anneal at about 450°C is done. This anneal, in a hydrogen ambient, is also done to anneal out interface traps at the Si-SiO_2 interface in the gates. Subsequent processing at about this same temperature can also occur, such as the deposition of an overlying dielectric or passivation layer.

However, the solubility of Si in Al is significant. At 450°C, the solubility is about 0.5 atomic percent, and at 500°C it is about 1 atomic percent. This means that pure Al in contact with Si will want to absorb the Si from the substrate up to its solubility level for that temperature. The only things limiting this process would be the diffusion of Si in the Al and the amount of Al acting as a sink for the Si. Unfortunately, the diffusivity of Si in polycrystalline Al is quite high, and the amount of Al nearby is large, so a significant amount of Si gets drawn up into the Al. This creates voids in the remaining Si, which are quickly filled by the overlying Al. If the Al penetrates uniformly into the substrate, typical processing conditions and contact/interconnect dimensions would result in about 0.2–0.3 μm of penetration. However, the situation is actually worse than that. That is because when the Al reduces the native oxide, it does so nonuniformly. Therefore the Al penetrates in localized spots. Al "spikes" occur in the Si substrate, penetrating in excess of 1 μm deep. Figure 11–8 illustrates this, showing the Al penetrating in some areas deeper than the diffused junction. This can short the junction. As a practical matter, pure Al contacts to Si cannot be used for junction depths of less than 2–3 μm.

One solution to the Al spiking problem is to use Al films that already have Si in them. Therefore the solubility requirement is already fulfilled. So Al films with approximately 1 atomic percent Si have been commonly used. This is the solubility of Si in Al at about 500°C, the highest temperature the Al sees during processing. The spiking problem is reduced, but another problem arises. When the contact structure is cooled down below

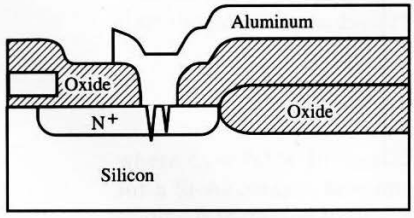

Figure 11–8 Si-Al contact region showing spiking of Al into Si active region. This is due to Si diffusing into the Al to satisfy the solubility requirement, with Al filling the resulting voids in the Si substrate.

500°C, or subsequent thermal steps are done at less than 500°C, the solubility of Si in Al is less. The now-excess Si precipitates out, usually at the interface where heterogeneous nucleation can occur, leaving Si nodules. In fact, there is enough Al dissolved in these nodules to make them p-type (Al is a p-type dopant in silicon), causing higher contact resistance to N^+ regions.

The use of Al(Si) to solve the spiking problem was adequate as long as the contact resistance was not limiting circuit performance. But as contact areas were reduced in size and the contact resistance did become an issue, another solution to Al-Si interaction and spiking was needed for contacts to small junctions. Barrier layers provided the answer. These layers are placed between the Al and Si to prevent or at least slow down any interaction. Such barriers need to be a barrier for chemical interdiffusion between Si and Al at processing temperatures (up to 450–500°C). They also need to be thermally stable. They should have low stress. This usually means that their coefficient of thermal expansion needs to be close to Si (2.6×10^{-6} °C^{-1}). They need to adhere well to Si and Al, as well as SiO_2. This usually means that there is some interfacial reaction between the barrier and Si and Al. Finally, they should have good electrical conductivity and a low contact resistance to both Si and Al.

Barrier layers have been classified by Nicolet [11.12] into three types: passive barriers, stuffed barriers, and sacrificial barriers. The passive barriers would be chemically inert to both Si and Al and be good diffusion barriers as well. TiN is an example of this type. Some materials are chemically inert, but diffusion along their grain boundaries can be quite significant. However, it has been found that by "stuffing" the grain boundaries with other atoms or molecules, the diffusion of other species can be greatly reduced, making them diffusion barriers as well. The reduction of diffusion by stuffing the grain boundaries is a combination of a physical effect (physically blocking the diffusing species) and a chemical effect (the stuffed species, such as nitrogen or oxygen, chemically bond with the diffusing species, such as Al, stopping their diffusion). A Ti-W alloy with nitrogen stuffing the grain boundaries—achieved by sputtering the Ti-W in an N_2 ambient—is an example of a stuffed barrier. Sacrificial barriers are those in which the barrier material is "sacrificed" in order to prevent reaction between the Al and Si. The material, such as Ti, reacts with either the Al or Si or both, forming aluminides or silicides. But by doing so, it prevents the Al and Si from interacting since the reactions with the barrier material are favored both kinetically and thermodynamically over the reaction between the Al and Si. However, once the sacrificial barrier material is consumed, the barrier is gone, and so these types of barriers are only effective for limited processing temperatures and times.

One of the first barrier layers for contacts was platinum silicide. It is formed by depositing Pt on the wafer, and then annealing at low temperatures (250–350°C). Because the surface Si is consumed during the formation reaction, a new, clean interface is obtained, and very good electrical contact is made. It is also "self-aligned" because the silicide is only formed where the Si is exposed through the oxide contact opening (and the unreacted Pt is chemically removed from other areas). Therefore an extra mask is not needed and no alignment overlap is required. PtSi acts satisfactorily as a barrier between Al and Si, but only if subsequent processing is done at about 350°C or below, and junction depths underneath are greater than about 0.3 μm. That is because PtSi will

react with Al, forming Pt-Al intermetallic compounds and eventually allowing the Al to spike into the Si substrate. Once junction depths got shallower than 0.3 μm and processing temperatures got above 350°C, another barrier was necessary.

Refractory metals by themselves have had limited use as barriers. Ti in particular is attractive since it reduces and breaks through any native oxide and adheres well to both Si and SiO_2. It forms very low resistive contacts, silicon diffusion through it is slow, and therefore it prevents the Si dissolving in the Al. However, it reacts with Al to form $TiAl_3$ at low temperatures. (Above 500°C, Ti will also react with the Si to form $TiSi_2$.) It thus acts as a sacrificial barrier, acting as a barrier only until it is consumed, which happens rather quickly above 400°C. So while Ti is a good contact/adhesion layer, it is not a very good barrier layer. Ti is thus often used in conjunction with a barrier layer on top.

Refractory metal silicides, such as $TiSi_2$, have also been used as barrier layers in contacts. They will also react with Al, but at a somewhat higher temperature (about 500°C for $TiSi_2$) than the noble metal silicides, such as PtSi. However, diffusion along their grain boundaries limits their effectiveness as barriers to below 400°C. Therefore, like Ti, refractory metal silicides are now usually used in contact structures as the adhesion/contact layer, with a better barrier layer on top. An advantage in using refractory metal silicides in contacts is the ability to form self-aligned contact structures. Ti can be deposited over both an oxide layer and into the openings in the contacts without a mask, but during an anneal it will only form a silicide in the contacts where Si is exposed. The unreacted Ti over the oxide can be etched off, leaving the silicide self-aligned to the contact.

In fact, the entire source and drain regions can be contacted by a refractory metal silicide this way, providing for local interconnects. By contacting the entire source and drain regions with silicide, rather than just in the contact holes, the contact resistance is also improved by providing a much larger contact area between the silicide and silicon. This is because the silicon/silicide contact resistance is much larger than silicide/barrier/metal contact resistances, and the net contact resistance thus depends on the silicon/silicide contact area.

As the backend processing temperatures were raised in later technology generations, the next commonly used barrier after PtSi was titanium-tungsten. This alloy, generally about 30 atomic percent, or 10 weight percent Ti, acts as a diffusion barrier between the Si and Al, especially when stuffed with nitrogen. While pure W is a good barrier itself, adding Ti helps it adhere to underlying layers and increases its corrosion resistance. Ti-W will react with Al at higher temperatures, about 600°C, while diffusion through the film limits its use to below 500°C. In addition, diffusion of Ti from the film into the Al at 400°C and above has been seen to raise the Al resistivity. Its contact resistance, however, is not as low as PtSi, possibly due to its difficulty in dissolving the native oxide on the Si surface. Thus a bilayer of PtSi and Ti-W has been generally used, taking advantage of the low contact resistance of the PtSi and the barrier properties of Ti-W. One problem with Ti-W is its relatively high stress. It is often very brittle, leading to cracking and peeling on the wafer and to flaking from chamber walls onto the wafer. [11.13] In addition, there is the problem of diffusion of Ti from the film into Al at 400°C and above, raising the resistivity of the Al interconnect line.

Refractory metal compounds, notably TiN, have become very popular barrier layers. These are commonly fine-grained structures, with grain sizes below 10 nm. Because of this, diffusion through them is very slow, making them impermeable to silicon and most other species. Incorporation of oxygen or nitrogen in these films makes them even better barriers. They are also chemically stable and chemically inert with most other layers. There may be some reaction with Al, forming some AlN and $TiAl_3$ at temperatures greater than 550°C. The electrical resistivity of TiN is low enough for use as both a contact material and a local interconnect. The contact resistance to Si is somewhat higher than that of PtSi, Ti, or $TiSi_2$. Therefore, TiN is usually used in a bilayer structure with TiN on top of Ti, with the Ti usually reacted to form $TiSi_2$ for a better contact. The best properties of both layers are utilized: the good adhesion and low contact resistance of the Ti or $TiSi_2$, and the good barrier properties of TiN. This is accomplished without much more processing, since the TiN can easily be deposited immediately after the Ti, just by adding nitrogen in the sputter deposition chamber. Or TiN can be formed on top of $TiSi_2$ by depositing Ti on top of the Si substrate in the contact opening, and annealing in a nitrogen ambient. The Ti reacts with the Si below it to form the silicide while simultaneously reacting with the nitrogen in the ambient to form TiN on top. This was in fact the process described in the CMOS process flow in Chapter 2 (Figures 2–3 to 2–33). The TiN can also be deposited by CVD.

Figure 11–9 shows a multilayer contact structure with $TiSi_2$/TiN contact/barrier layers. The $TiSi_2$ makes good electrical contact to the Si and provides good adhesion. The TiN serves as a barrier to interaction between the Al and both the $TiSi_2$ and Si. Table 11–3 lists commonly used contact barrier layers and compares their maximum processing temperature and modes of failure. It is clear that early contact structures have evolved from simple Al-on-Si structures to barrier-layer structures to multilayer structures. We will later see how W plugs are often used in contacts to improve step coverage and planarization.

11.2.2 Interconnects and Vias

Interconnects have similarly evolved as the technology has progressed and the requirements placed on interconnects have increased. The primary requirements on interconnects are low resistivity, good adhesion to underlying films, stability during processing and operation, and the ability to be deposited and etched.

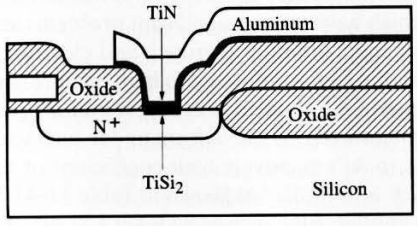

Figure 11–9 Multilayer contact structure with $TiSi_2$/TiN contact/barrier layers between the Si and Al.

Table 11–3 Barrier-layer properties [11.14, 11.15]

Structure	Failure Temperature (°C)	Failure Mechanism (Reaction products)
Al/PtSi/Si	350–400	Compound formation (Al_2Pt, Si)
Al/$TiSi_2$/Si	400	Diffusion ($Al_5Ti_7Si_{12}$, Si at 550°C)
Al/NiSi/Si	400	Compound formation (Al_3Ni, Si)
Al/$CoSi_2$/Si	400	Compound formation Al_9Co_2, Si)
Al/Ti/PtSi/Si	450	Compound formation (Al_3Ti)
Al/$Ti_{30}W_{70}$/PtSi/Si	500	Diffusion (Al_2Pt, $Al_{12}W$ at 500°C)
Al/TiN/$TiSi_2$/Si	550	Compound formation (AlN, Al_3Ti)

Aluminum, aluminum alloys, and aluminum containing multilayers have been the dominant interconnects in silicon technology to this point in time. Al has low resistivity (although not low enough to meet future NTRS specifications), adheres well to Si and SiO_2, can reduce other oxides, and can be etched and deposited using reasonable techniques. Its main shortcoming is its stability during processing and operation, which we will discuss. For early MOS technology, Al was used not only as the primary interconnect (i.e. global) but also as the gate electrode and a local interconnect. However the melting temperature of Al is low (660°C for pure Al), and it cannot withstand high processing temperatures. When self-aligned source and drain regions were needed to reduce gate-source and gate-drain capacitances, a new material was needed for the gate electrode that could withstand the 800–1000°C diffusion steps or implant anneals needed to dope the source and drain regions while the gate electrode material was present. Heavily doped polycrystalline silicon was the solution to this problem. Polysilicon can be utilized as an interconnect as well, although only as a local interconnect because of its resistivity.

The first modification to pure Al interconnects was the addition of Si at about 1 atomic percent. This prevents spiking of Al into the silicon substrate which is caused by silicon's relatively high solubility and diffusivity in Al. Even with the use of barrier layers such as TiN, Si is often still used in Al interconnects. This is because the presence of the Si reduces the driving force for interdiffusion and hence decreases the chance of barrier failure. Si also suppresses the reaction between Al and TiN.

Almost all the rest of the modifications made to the Al interconnect structure have been as a result of aluminum's low melting point and ease of deformation. Two phenomena that occur with aluminum which have caused significant problems and have required modifications in the Al interconnect are hillock growth and electromigration.

Hillock growth occurs when the Al thin film is subjected to high compressive stresses. Because of the polycrystalline structure of Al, coupled with its low melting point and resulting susceptibility to plastic deformation, the Al will deform under such stresses. High stresses can easily be generated due to Al's relatively high coefficient of thermal expansion (23×10^{-6} °C^{-1}, while Si is 2.6×10^{-6} °C^{-1} as shown in Table 11–4). If Al is deposited on a Si substrate (or even on another film, such as SiO_2, on a Si substrate) at low

Table 11–4 Mechanical properties of interconnect materials

Material	Thermal expansion coefficient (°C^{-1})	Elastic modulus, $Y/(1-\nu)$ (MPa)	Hardness (kg mm^{-2})	Melting point (°C)
Al (111)	23.1×10^{-6}	1.143×10^5	19–22	660
Ti	8.41×10^{-6}	1.699×10^5	81–143	1660
$TiAl_3$	12.3×10^{-6}	—	660—750	1340
Si (100)	2.6×10^{-6}	1.805×10^5	—	1417
Si (111)	2.6×10^{-6}	2.290×10^5	—	1417
SiO_2	0.55×10^{-6}	0.83×10^5	—	~1700

temperatures and the structure is heated, high compressive stresses in the Al can be generated. This is because the Al, which is tightly constrained to the wafer, will want to expand much more than the Si. To relieve the stress, portions of the Al are squeezed up, forming small hills or "hillocks." The movement of the Al occurs primarily along grain boundaries in the Al since diffusion is usually much faster along grain boundaries than through the interior of the grains. Whole grains of Al are often pushed or grow upward, forming hillocks about the size of the grains, 0.5 to 3 μm tall and wide.

Figure 11–10 shows schematically the process of hillock formation. Hillocks can result in electrical shorts between interconnect levels, as well as cause the surface topography to be rough, making lithography and etching more difficult. While the overlying films can help suppress hillock formation by acting as a mechanical barrier, this does not completely eliminate the problem. In fact, so much stress can build up that an overlying SiO_2 layer can crack. When this happens, the Al can protrude from the crack in long whiskers many microns long, increasing the chances of a short.

The addition of elements that have limited solubility in Al, such as Cu, has been found to suppress hillock formation. The excess atoms segregate and precipitate preferentially at the Al grain boundaries. This reduces the Al diffusion along the grain boundaries and suppresses hillock formation.

In a similar manner, voids can form when the Al is subjected to tensile stresses. Upon cooling, the Al will want to shrink more than the Si below it. (Unfortunately any hillocks that have formed during the previous heating do not significantly reverse themselves and disappear during the cooling.) Since the Si is more mechanically rigid than the Al, the stress is relieved by vacancy movement and agglomeration in the Al, with voids forming. This can greatly increase the interconnect electrical resistance and even cause open circuits. The addition of Cu to the Al can suppress void formation in Al as well as suppressing hillock formation.

Perhaps the biggest problem encountered with Al interconnects is that of electromigration. One of the requirements of an interconnect is that it be stable not only during processing, but also during circuit operation. Electromigration is a phenomenon that occurs during circuit operation and has very adverse effects. When an electrical current flows through aluminum (on the order of 0.1–0.5 MA cm^{-2}, which is commonly achieved in integrated circuits), the electrons can actually transfer enough momentum

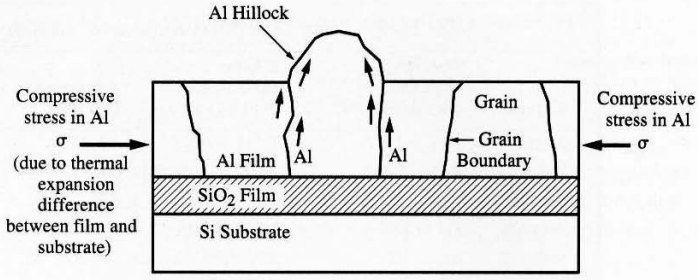

Figure 11–10 Schematic illustration of hillock formation due to compressive stress in an Al film. Al diffusion along grain boundaries is indicated.

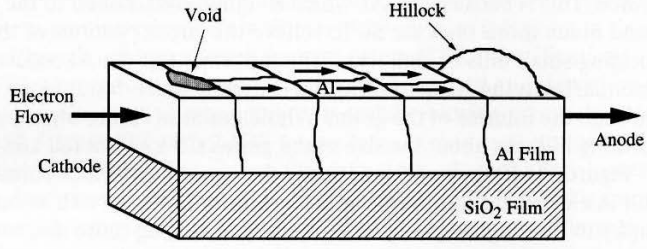

Figure 11–11 Schematic illustration of electromigration with resultant hillock and void formation. Diffusion of Al through grain boundaries is indicated.

to the Al atoms to cause them to diffuse. This diffusion is faster along grain boundaries, and it can cause a buildup of Al in some regions, resulting in hillocks, and can cause a depletion in other regions, resulting in voids. As a result shorts and opens in the interconnects can occur during the normal operation of the circuit. Figure 11–11 shows this process schematically, and Figure 11–12 shows a SEM of an actual circuit where this has occurred. Because most of the Al atoms and vacancies diffuse along the grain boundaries, electromigration is very dependent on the grain structure, including the size and crystallographic orientation of the grains. It also depends on over- and underlying layers, and on the processing history of the interconnects.

Substantial effort has been made to study and remedy this phenomenon. One of the early remedies for electromigration was to add Cu to the Al. As is the case in suppressing hillocks, adding Cu is believed to suppress grain boundary diffusion and hence slow down electromigration. Adding too much Cu, however, causes problems with etching of the interconnect as well as corrosion, and 4 weight percent Cu is the most that is usually added.

Thus, the first modifications to Al interconnects involved switching from pure Al to Al with about 1–2 weight percent Si and 0.5–4 weight percent Cu. Adding these elements to the Al increases the sheet resistivity of the interconnect by as much as 35%,

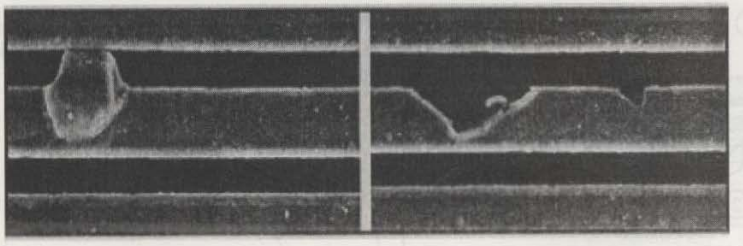

Figure 11–12 SEM top view of hillock and voids that have formed due to electromigration in an Al line [11.16]. Reprinted by permission of Materials Research Society.

but this was a necessary trade-off in order to prevent the failure of circuits due to electromigration and hillock and void formation. Other solutions, similar to those for contacts involving multilayer structures, were later developed and will be discussed later.

The next change to interconnects involved the local interconnects. Polysilicon has been used as the gate electrode since the mid 1970s. But as circuits got faster, the sheet resistivity of heavily doped polysilicon became a concern (Figure 11–4). In addition, the doped silicon in the source and drain regions has a fairly high resistivity (50–100 $\mu\Omega$-cm) which can add appreciable parasitic resistance to MOS transistors. Silicide materials can be used to reduce all these problems.

Silicides have been utilized as local interconnects in three primary ways, as illustrated in Figure 11–13. On top of the polysilicon layer they form a "polycide" structure; on top of the source and drain regions they strap the doped silicon; and finally, they can be used as an independent local interconnect. All of these uses were illustrated in the CMOS process in Chapter 2 (Figure 2–33). In the polycide structure, the silicide decreases the sheet resistance of the polysilicon gate and local interconnect, while retaining the highly reliable polysilicon/SiO_2 interface. On top of the source and drain regions, the silicide reduces the resistance of the diffused layers. The silicide layer also provides a good contact to the silicon in the gate, source, and drain regions, while increasing the contact area. Silicides are also sometimes used as local interconnects by themselves, providing relatively low resistive paths below the lowest metal layers. Silicides are used because they offer low electrical resistivity (10–50 $\mu\Omega$-cm), they are stable at high temperatures, they provide good process compatibility with silicon, they are easy to plasma etch, they provide good contacts to other materials, and they do not exhibit much electromigration. Some of the common silicides used in silicon technology are listed in Table 11–5, along with some of their properties.

As discussed in Chapter 9, the formation of silicides can be accomplished in two ways: by direct deposition of the silicide and by the reaction between the metal and Si on the wafer to form the silicide. The direct deposition method can use sputtering from a composite target, cosputtering from two targets of the metal and Si, co-evaporation of the metal and Si, or Chemical Vapor Deposition (CVD). CVD is currently only done for WSi_2. The reaction method can be accomplished by two different techniques. The first technique is by depositing the metal and amorphous Si, both usually by sputtering, and

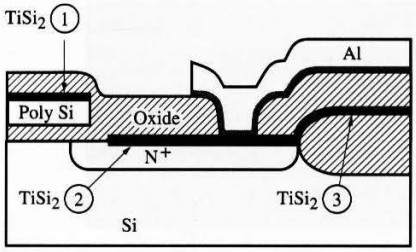

Figure 11–13 Uses of silicides in back-end structures: 1, on top of a polysilicon gate; 2, on top of source and drain regions; and 3, as a local interconnect.

Table 11–5 Properties of silicides (After [11.17])

Silicide	Thin film resistivity (μΩ-cm)	Sintering temp (°C)	Stable on Si up to (°C)	Reaction with Al at (°C)	nm of Si consumed per nm of metal	nm of resulting silicide per nm of metal	Barrier height to n-Si (eV)
PtSi	28–35	250–400	~750	250	1.12	1.97	0.84
$TiSi_2$ (C54)	13–16	700–900	~900	450	2.27	2.51	0.58
$TiSi_2$ (C49)	60–70	500–700			2.27	2.51	
WSi_2	30–70	1000	~1000	500	2.53	2.58	0.67
Co_2Si	~70	300–500			0.91	1.47	
CoSi	100–150	400–600			1.82	2.02	
$CoSi_2$	14–20	600–800	~950	400	3.64	3.52	0.65
NiSi	14–20	400–600	~650		1.83	2.34	
$NiSi_2$	40–50	600–800			3.65	3.63	0.66
$MoSi_2$	40–100	800–1000	~1000	500	2.56	2.59	0.64
$TaSi_2$	35–55	800–1000	~1000	500	2.21	2.41	0.59

patterning and selectively etching the amorphous Si. (The Si may be deposited and patterned either before or after the metal deposition.) The wafer is then annealed so that the metal and amorphous Si react to form the silicide, and the unreacted metal is selectively etched off. The second reaction method is similar but no amorphous Si is deposited. Instead, the metal is deposited on the exposed gate and/or source drain regions, all of which are silicon. The wafer is annealed and the silicide forming reaction occurs wherever the silicon and metal are in contact, that is, on the gate and source/drain regions. The unreacted metal is then selectively etched away. The first of these reaction techniques produces local interconnects independent of either gate or source/drain regions, while the second reaction method produces self-aligned silicide structures on the gate and/or source drains. This second technique is often called a salicide, for "self-aligned silicide." Figure 11–14 illustrates the salicide process.

A high-temperature anneal is required for either the direct deposition or reaction methods of silicide formation. In the reaction method, a high-temperature step is needed to react the metal with the Si to form the silicide and hence to achieve low re-

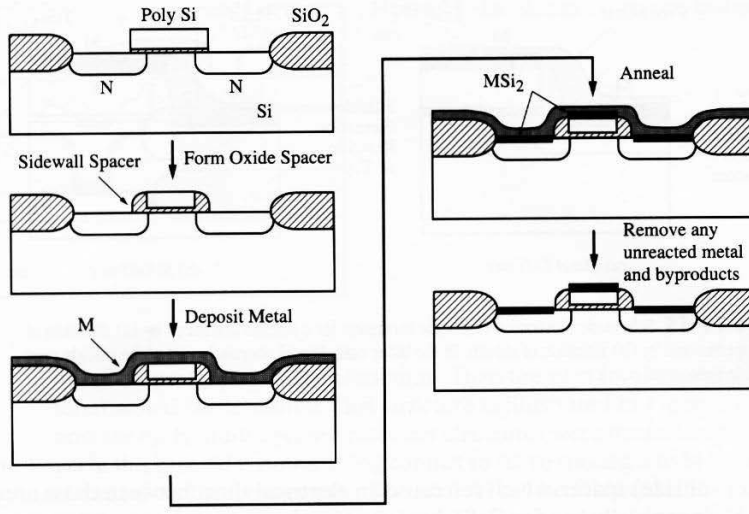

Figure 11–14 Salicide (self-aligned silicide) process. M represents a metal, such as Ti.

sistivity films. A high-temperature step is also required in the direct deposit methods. Depending on the exact deposition technique, the grain structure of the as-deposited film is either very fine, microcrystalline, amorphous, or even an unreacted multilayer structure. In any of these cases, annealing is needed to make the grains grow in size in order to achieve low-resistivity films. This is because in small-grain films there are more grain boundaries which scatter the carriers, reducing the electrical mobility. The final, or minimum, resistivity for various silicides is given in Table 11–5.

There have been several silicides used in silicon technology, including WSi_2, $TaSi_2$, $MoSi_2$, $TiSi_2$, and $CoSi_2$. Tungsten silicide has been used because a CVD deposition process is available for it, providing good step coverage and good process control. In addition, it can be oxidized rather easily, as long as a source of silicon is maintained below it. Therefore, a good thermal oxide can be formed over it rather than a deposited oxide. However, high stresses in WSi_2 often cause stability and adhesion problems. WSi_2, as well as $TaSi_2$ and $MoSi_2$, are not as easy to form through reaction with Si as $TiSi_2$ and $CoSi_2$ are, and therefore are not as attractive for salicide applications.

$TiSi_2$ is very popular due to its low resistivity, and its ability to reduce native oxides and to adhere well to most other materials. However, $TiSi_2$ has a high-resistivity phase ("C49") that forms at moderate temperatures. This must be transformed to a low resistivity phase by annealing, and this transformation becomes more difficult for narrow and thin lines. Also, the thin film has a tendency to agglomerate, or roughen up, at higher annealing temperatures.

$CoSi_2$ behaves better in regard to these particular problems. In addition, it generally shows less lateral encroachment during its formation reaction. Lateral encroachment means the silicide forms beyond the source, drain, and gate regions over the oxide (or

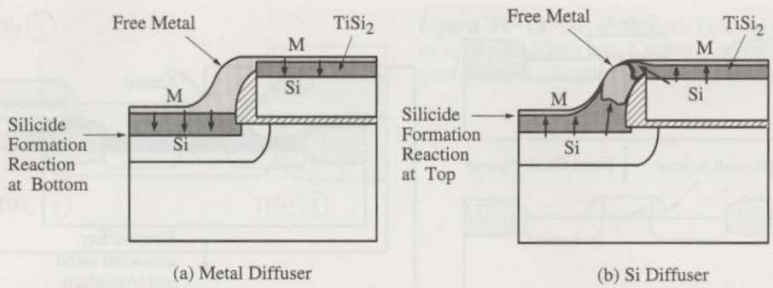

Figure 11–15 Schematic illustration of silicide formation for a silicide that forms by (a) diffusion of metal species and by (b) diffusion of silicon. In the latter case, lateral encroachment of the silicide over the oxide spacer can occur.

nitride) spacer, which can cause an electrical short between those areas. Encroachment is generally less for $CoSi_2$ because of its lower reaction temperatures, and also because the metal is believed to be the dominant diffusing species during silicide formation. This is shown schematically in Figure 11–15. However, the resistivity of $CoSi_2$ is slightly higher than $TiSi_2$, and it does not have as good adhesion and oxide reducing properties as $TiSi_2$. $CoSi_2$ is more sensitive to surface contaminants and oxides, which can lead to non-uniform silicidation and junction leakage.

TiN, as a byproduct of the $TiSi_2$ salicide process, can also be utilized as a local interconnect as was illustrated in the CMOS process flow in Chapter 2 (Figures 2–31 to 2–33). The reaction between Ti and Si in the salicide process is often done in a nitrogen ambient in a two-step anneal process, as will be discussed in the next section on processing techniques. Portions of the TiN that is formed can be masked and left on the circuit to act as a local interconnect. An advantage of using TiN as a local interconnect rather than polysilicon or a silicide is that TiN is a diffusion barrier for most dopants. TiN is commonly utilized as a local interconnect in SRAM cells, where small cell size is critical, connecting load drains with the input gate of the inverter.

Another material that has been used or considered for a local interconnect is tungsten. Selective deposition of tungsten over Si areas using CVD (see Chapter 9) has been used to form W/polysilicon gate and W/silicon source and drain structures. However, at temperatures above 600°C the W and silicon will react to form a silicide, eliminating the advantage of having a lower resistivity material strapping the gate, source, and drain regions. First-level dielectric processing is commonly done above this temperature, which would be a roadblock to utilizing W in this manner.

As circuits became larger and more complex, additional interconnect levels have been required (Table 11–1). Adding an additional level of global interconnect seems easy enough. The process could be as simple as opening a hole in the top dielectric level wherever a contact to the lower level is needed, depositing Al (with some Cu in it to inhibit electromigration), forming both the contact and the next level interconnect, fol-

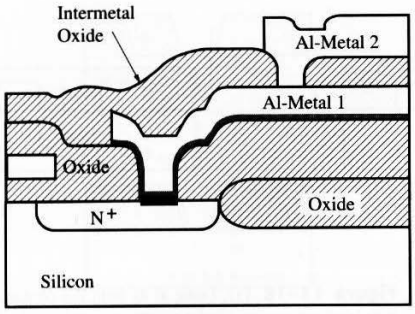

Figure 11–16 Illustration of early two-level metal structure.

lowed by patterning of the aluminum. Then the next level of intermetal dielectric material would be deposited. This structure is illustrated in Figure 11–16. This, in fact, is how the early multilevel interconnect structures were made. Looking at the contact, or via in this case, Al is now making contact to Al, so spiking into Si is not a problem. Good electrical contact between two layers of Al can be achieved if the interface is clean.

However, the topography gets less planar as more levels are added, as Figure 11–16 illustrates. A more planar topography is desired for two main reasons—lithography limitations and step coverage. The lithography limitations arise partly from the limited depth of focus of submicron lithography instruments. The maximum depth of focus is about ± 0.35 μm for state-of-the-art optical tools (see Chapter 5.). In addition, nonplanar topography results in nonuniform resist thickness, which can result in nonuniform exposure and development. Another lithography issue is that spurious UV light reflections can occur from the sides of metal as it goes over steep steps, again causing lithography problems. There are lithographic techniques that address these problems, such as using multilayer resist structures as discussed in Chapter 5, but these are not always effective or practical solutions.

Nonplanar topography also leads to problems with step coverage and filling of spaces with deposited films. We saw in Chapter 9 how step coverage and filling can be problems, especially for PVD deposited metal (see Figure 9–1, for example). Such problems lead to higher electrical resistance in the line, as well as potential reliability problems. To make matters worse, step coverage problems compound one another. A large step height on one level can lead not only to thinning of the next level on sidewalls, but also cusping and overhangs. This in turn leads to even worse coverage problems on the next level.

Reducing the step heights and achieving more planar topology through processing techniques is called planarization. The Degree Of Planarization (DOP) is defined as

$$\mathrm{DOP} = 1 - \frac{x^f_{\mathrm{step}}}{x^i_{\mathrm{step}}} \tag{11.13}$$

where x^i_{step} is the initial step height in the topography and x^f_{step} is the final step height after some sort of planarization is done. These step heights can refer to hole depths as

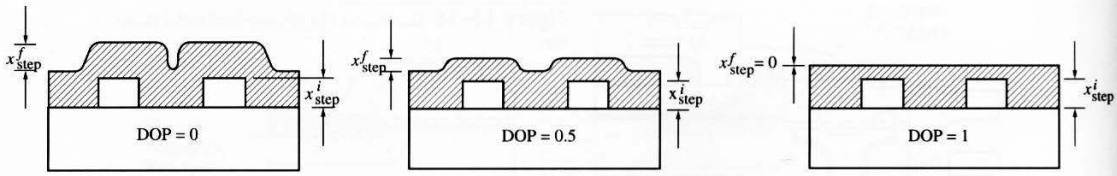

Figure 11–17 Demonstration of Degree Of Planarization (DOP).

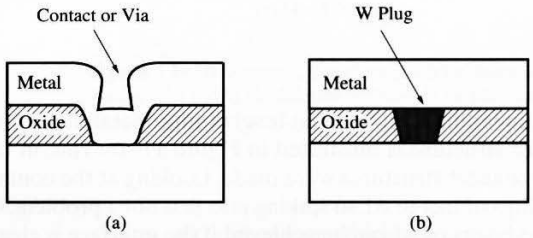

Figure 11–18 (a). Using Al as both the contact and the next level interconnect layer results in nonplanar topography. (b) Using a tungsten plug as the contact results in planar topography.

well. A DOP of 0 means that no planarization is achieved, while a DOP of 1 corresponds to complete planarization—the final step height is zero. Figure 11–17 shows different degrees of planarization. Several methods have been developed to improve the planarity of multilevel interconnect structures. Some involve the metals, which will now be discussed, while others involve the dielectric materials, which will be discussed later.

One way to help achieve planar metal layers is by using a W contact or via fill, also called a W plug. Figure 11–18(a) shows the topography above a contact or via which uses Al as both the contact material and the next level of interconnect. Not only might the step coverage of the Al be a problem, but the resulting topography over the contact can be very non-planar. Using a via fill or plug structure results in the structure in Figure 11–18(b). This can be accomplished in two ways. One method involves a selective W deposition into contact or via holes cut in the dielectric layer, which fills the contacts or vias without depositing any W over the top of the dielectric. The other, more common, method is using a blanket W deposition, which fills the contact or via as well as covering the dielectric. Then an "etchback" is done to remove the W over the top surface of the dielectric. This is illustrated in Figure 11–19. This process of filling holes or trenches in a dielectric level with the metal (usually followed by an etchback), rather than depositing, patterning, and etching the metal on top of a dielectric layer, is called the "damascene" method. Its name comes from the ancient practice of inlaying metal in wood or ceramics for decoration. It can be used for vias or interconnect layers, or both.

Using a W plug produces a much smoother topography over the contact or via, as shown in Figure 11–18(b), resulting in a near-planar contact or via level. The DOP approaches 1 here, where the initial step height corresponds to the height of the contact or via hole. Figure 11–2 illustrated such W contact plugs with an Al metal line on top. With a planar contact or via level, the problem with step coverage of sputtered Al is greatly reduced since there are few, if any, steps for the Al to cover. W is used for the

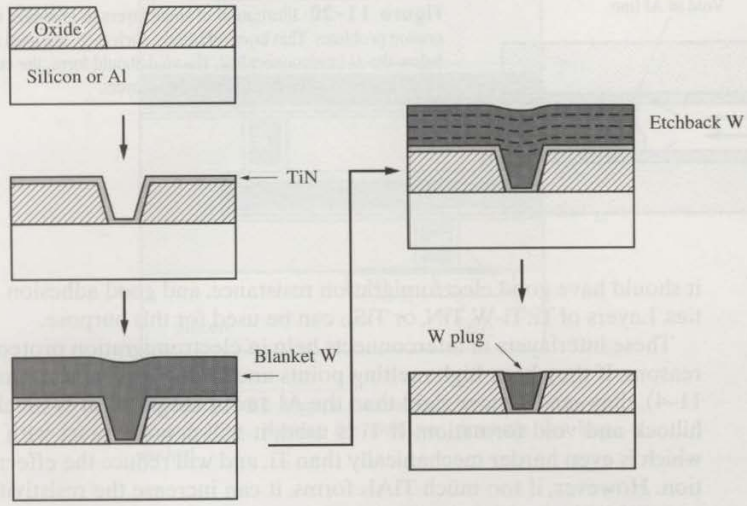

Figure 11–19 Schematic illustration of process to fabricate a W plug for a contact or via. TiN (or a Ti/TiN bilayer) is utilized as an adhesion and barrier layer. An etchback step is done to planarize the W and leave it only in the contact or via hole.

plug rather than Al because the deposition of W using the CVD technique gives much more conformal filling of the contact or via hole. Because of problems with CVD deposition of Al-Cu alloys, as discussed in Chapter 9, CVD Al processes are generally not used in manufacturing.

One planarizing technique under development that does utilize Al as the contact or via metal is called Al reflow. By heating the Al to 450–550°C either during the deposition or immediately afterward, the Al in a contact or via hole will flow and fill the hole, as well as help planarize itself. This is similar to reflow of dielectric materials which are heavily doped with B or P. With this technique, one could deposit the via (or contact) and interconnect levels in one step to obtain good filling of the contact and good step coverage and planarization of the interconnect level.

Just as contacts benefited from the use of multilayer structures, so have interconnects. This is due to a variety of reasons, most of them related to issues we have already discussed. First, electromigration has continued to be a problem that has hampered the reliability of silicon integrated circuits. While the addition of Cu to the Al has helped, it has not completely eliminated the problem. Other methods have been used to inhibit electromigration, or in some cases, negate its effect on the integrity of the interconnect lines. One method uses shunt lines. Another interconnect material is layered below, above, or within the Al layer, as shown in Figure 11–20. If electromigration does cause an opening in the Al line, the current can be shunted through the other material, maintaining the electrical integrity of that interconnect. The material does not have to have as low a resistivity as Al, since current will only flow through it for a short distance. But

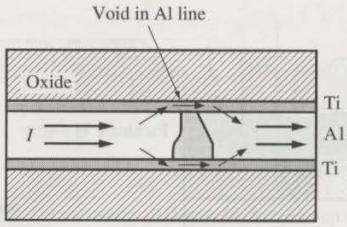

Figure 11–20 Illustration of shunt layers to minimize electromigration problems. Thin layers of metals, such as Ti, are used above and below the Al interconnect line. If a void should form, the current can be shunted around it and through these layers.

it should have good electromigration resistance, and good adhesion or barrier properties. Layers of Ti, Ti-W, TiN, or $TiSi_2$ can be used for this purpose.

These interlayers in interconnects help in electromigration protection for two other reasons. If they have high melting points and strong mechanical properties (see Table 11–4), they can be more rigid than the Al and act as physical or mechanical barriers to hillock and void formation. If Ti is used, it will actually react with Al to form $TiAl_3$, which is even harder mechanically than Ti, and will reduce the effects of electromigration. However, if too much $TiAl_3$ forms, it can increase the resistivity of the interconnect. If TiN is used, it is best to have excess Ti in it which will react with the Al to form $TiAl_3$. In addition, the electromigration of Al is very dependent on the grain texture of Al. The closer all the grains are to a single, usually (111), orientation in the vertical direction, the less electromigration occurs. The grain texture of the Al is very dependent on what underlayer it is deposited on. Al deposited on Ti exhibits much better grain texture (more grains closer to a single orientation) and electromigration behavior than on other films, such as Si, SiO_2, or TiN. If TiN is used between the Ti and Al, as a reaction barrier for example, the Al still has good grain texture. This is because the TiN takes on a pseudogranular structure, passing the Ti structure to the Al interface. Therefore, by using either Ti or Ti/TiN layers underneath the Al layers, electromigration is decreased due to grain texture considerations.

Interconnect and via structures made up of multilayers have been developed for barrier and adhesion reasons, just as with the contact structures. Ti can help with adhesion and electrical contact. TiN can serve not only as a good barrier, but as an adhesion layer as well, particularly in conjunction with W plugs. A TiN or Ti/TiN layer provides for good contact between Al and W and prevents any reaction between them. In addition, when CVD W is deposited in a via, a barrier layer must be used between the W and the underlying Al or else the fluorine in the WF_6 used in the CVD W deposition process will react with the exposed Al. (Similarly, in a contact a barrier is required to prevent the WF_6 from reacting with Ti or $TiSi_2$ layers or with Si.) Since the adhesion of W to dielectric layers is very poor, an adhesion layer is required and TiN (or Ti/TiN) is commonly used for this. A TiN layer can also be used in the interconnect stack to serve as a barrier between the Ti and the Al.

Another reason for using a multilayer interconnect structure is for lithographic purposes. The top Al surface is very reflective, and the reflection of light can adversely and unevenly affect the lithographic process. Therefore, an antireflective material is often deposited on top of the Al to minimize this. TiN again is usually used.

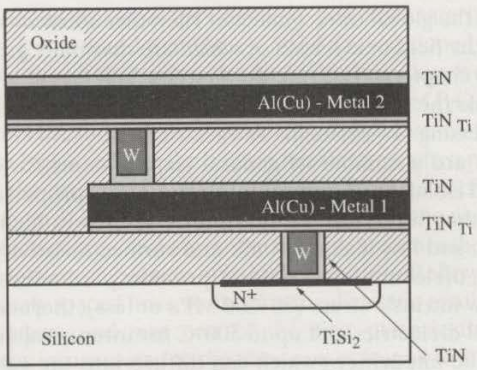

Figure 11–21 A typical metalization scheme used in current IC technology, showing the multilayer structures that have evolved.

For all these reasons, multilayer interconnect structures have been developed. One such structure is illustrated in Figure 11–21 and serves to summarize the issues we have discussed. A W plug is used in both the contact and the vias, resulting in good planarity. At the bottom of the contact, $TiSi_2$ is used to make good electrical contact to the Si, and also serves to strap the sources and drains, providing for lower sheet resistance along those paths. In both the contacts and the vias, TiN surrounds the W, promoting adhesion to the oxide. It also acts as a barrier to prevent attack of the underlying material by WF_6 during W deposition. The global interconnect structure, or "stack," in this example is Ti/TiN/Al(Cu)/TiN. Cu in the Al suppresses electromigration and hillock formation. Ti improves the grain texture of the Al for better electromigration protection and promotes adhesion. TiN is a barrier between the Al and Ti to limit $TiAl_3$ formation. The Ti/TiN and TiN layers on the bottom and top of each interconnect provide an electrical shunt in case of void formation in the Al. They also suppress hillock and void formation by providing mechanical barriers to plastic deformation. TiN on the top also acts as an antireflective layer for lithography. Many variations of this structure are used by different manufacturers, especially in the exact makeup and order of materials in the interconnect stack (the Ti or TiN is often omitted from the bottom of the stack, for example). But the basic idea of multilayer planar structures is very common.

11.2.3 Dielectrics

Dielectric layers are used to separate and electrically isolate conducting layers from each other. Two types of dielectric layers associated with metallization schemes were shown in Figure 11–1. The first-level dielectric is a deposited layer that separates the global (usually Al) interconnects from the diffused substrate areas, polysilicon, and local interconnects. The intermetal dielectrics (IMDs) are deposited dielectrics that

separate the global interconnects. (The other common dielectric layers are the gate oxide and the field oxide, both of which are thermally grown and have been discussed in previous chapters.) Because the first-level dielectric is deposited before any Al is present, while the intermetal dielectrics are deposited after Al is present, they have different processing requirements and restrictions.

There are a number of general requirements for dielectric layers [11.18]. Good electrical isolation is paramount and includes properties such as a low dielectric constant so that interconnect capacitances are low; a high breakdown field strength (>5 MV/cm); and low leakage (bulk and surface resistivity should exceed 10^{15} Ω cm). In addition, dielectric layers should provide good adhesion to silicon, metals, and silicides; low intrinsic stress (50–100 MPa or less); thermal stability (up to 850–950°C for first-level dielectrics and up to 500°C for intermetal dielectrics); contain no moisture or metallic impurities (which can diffuse into the substrate and cause leakage current); low-defect density (pinholes and particulates); high-structural density (low porosity); contain no residual constituents that can outgas; and they should be permeable to hydrogen (so that passivation of the Si/SiO_2 interface states can take place). Table 11–6 lists several dielectric materials used or being considered for use in back-end structures.

Table 11–6 Materials used or considered for interlevel dielectrics (After [11.19])

Material class	Material	Dielectric constant	Deposition technique
Inorganic	SiO_2 (including PSG and BPSG)	3.9–5.0	CVD Thermal oxidation Bias-sputtering High-density plasma
	Spin-on-glass (SiO_2) (including PSG and BPSG)	3.9–5.0	SOD (spin-on-dielectric)
	Modified SiO_2 (e.g., fluorinated SiO_2 or hydrogen silsesquioxane—HSQ)	2.8–3.8	CVD/SOD
	BN (Si)	>2.9	CVD
	Si_3N_4 (only used in multilayer structure)	5.8–6.1	CVD
Organic	Polyimides	2.9–3.9	SOD/CVD
	Fluorinated polyimides	2.3–2.8	SOD/CVD
	Fluoro-polymers	1.8–2.2	SOD/CVD
	F-doped amorphous C	2.0–2.5	CVD
Inorganic/Organic Hybrids	Si-O-C hybrid polymers based on organosilsesquioxanes (e.g., MSQ)	2.0–3.8	SOD
Aerogels (Microporous)	Porous SiO_2 (with tiny free space regions)	1.2–1.8	SOD
Air bridge		1.0–1.2	

CVD silicon dioxide is the dominant dielectric material used in back-end structures. In the 1960s, it was found that if the CVD SiO_2 was doped with phosphorus, forming phosphosilicate glass, better properties for MOS technology were obtained (see Chapter 4, Section 4.2.3). Phosphosilicate glass (PSG) is made up of P_2O_5 and SiO_2. The phosphorus in the films allows the film to getter, or tie up, sodium ions, resulting in more stable MOS characteristics. It also reduces the built-in tensile stress often present in deposited SiO_2 films. In addition, the presence of phosphorus allows one to "reflow" the layer. By annealing the PSG in steam at a temperature from 950–1100°C, the dielectric layer flows in order to reduce surface energy, and any sharp corners are smoothed out. Undoped oxide does not flow nearly as much because of its higher viscosity. Reflowing also helps in filling spaces between closely spaced lines. Reflowing the dielectric results in better step coverage of subsequent layers and was one of the first planarization techniques. The amount of phosphorus incorporated in the films ranges from 4 to 8 wt. percent, with more phosphorus leading to more reflow. However, more phosphorus also enhances the film's absorption of water. Atmospheric moisture absorbed in the film can react with the phosphorus to form phosphoric acid, which can lead to corrosion of aluminum layers on top of and below the film. Absorbed water can also lower the dielectric constant of the film, and can outgas during heating, causing metal deposition problems. A sandwich structure is often used with layers of undoped SiO_2 below and above the PSG to help protect against these problems.

PSG has commonly been used as the first-level dielectric. Since Al is not present at this point in the processing, high-temperature densification and reflow can be used. If silicides are present, temperatures below 950°C are required for first dielectric-level processing since most silicides start to degrade at about 900°C. If boron is added to PSG, to form BPSG (borophosphosilicate glass), the reflow temperature is reduced to about 800°C. Reflowed BPSG is illustrated in Figure 11–22. BPSG, with 2–6 wt. percent boron and phosphorus, also traps sodium and reduces the stress in the film.

When two metal-level interconnect structures were first introduced, PSG or BPSG were initially used for the intermetal dielectric. But since the structures with Al layers could not be heated above 450°C, a reflow step could not be done, and these films had many problems. Because of the relatively high sticking coefficients (around 0.3 to 0.4) characteristic of silicon dioxide films formed by the oxidation of silane, conformal depositions were usually not achieved. This can lead to problems with step coverage and filling of spaces. These problems get worse as metal spacings shrink with each new

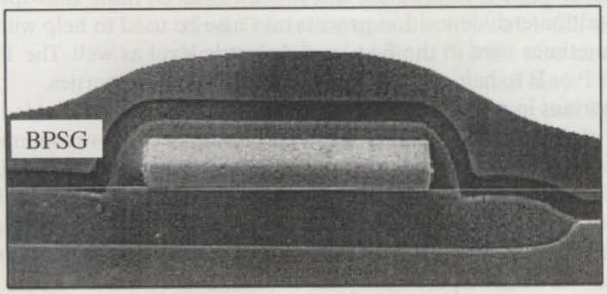

Figure 11–22 SEM image of BPSG oxide layer after 800°C reflow step, showing smooth topography over steps. Photo courtesy of VLSI Technology, Inc.

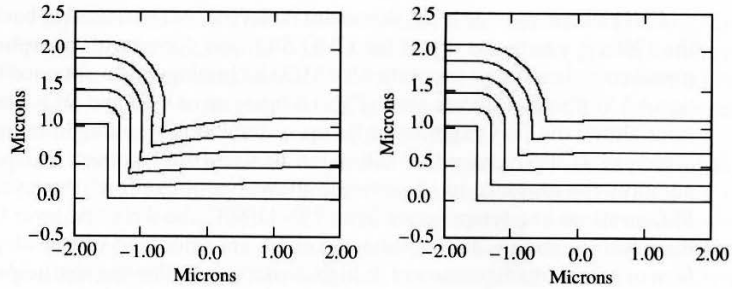

Figure 11–23 SPEEDIE simulations of silicon dioxide depositions over a step for silane deposition ($S_c = 0.4$) on left and TEOS deposition ($S_c = 0.1$) on right, showing less cusping in the latter case.

generation, especially if metal thicknesses do not shrink as fast. Without a reflow step to smooth out the films and fix the coverage and fill problems, failures in the interconnects can result. In addition, low-temperature SiO_2 films that are not annealed at higher temperatures are not as dense and not as good quality. Other deposition and processing techniques, as well as different materials, were needed for intermetal dielectrics.

Plasma-enhanced CVD techniques provide better oxide films for intermetal dielectrics. The energy from the plasma enhances the deposition reactions and provides for denser films and somewhat better coverage and filling at lower temperatures. Silicon dioxide films deposited by decomposing TEOS (tetraethooxysilane—$Si(OC_2H_5)_4$), rather than by reacting oxygen and silane, have also been developed. The reactant species in these depositions tend to have a low sticking coefficients, on the order of 0.1. This results in more conformal films than those produced from silane for both LPCVD and PECVD systems. Figure 11–23 shows SPEEDIE simulations of LPCVD silicon dioxide depositions over a step for silane deposition ($S_c = 0.4$) and TEOS deposition ($S_c = 0.1$). Less cusping occurs in the latter case, resulting in better conformality as well as filling ability. Even better conformality and filling ability at low temperatures is achieved by reacting the TEOS with ozone. Low-temperature deposition processes, using ozone or PECVD, allow the use of TEOS oxides as intermetal dielectrics. The non-plasma, TEOS/O_3 films are more porous than silane based or PECVD TEOS oxides and absorb more moisture. Often a thin layer of PECVD TEOS or silane oxide is used below and above TEOS/O_3 films. Bias-sputtering or a sequential deposition/etch/deposition process may also be used to help with filling. TEOS films are sometimes used in the first-level dielectric level as well. The TEOS films can be doped with P or B to help with Na gettering and reflow properties.

Planarization is important in multilevel back-end structures and can be divided into local planarization and global planarization. Local planarization occurs over submicron or micron (device) dimensions, while global planarization occurs over circuit or chip dimensions. Figure 11–24 illustrates this. Global planarization is usually desired, but local planarization is easier to achieve. An example of local planarization would be the reflow process, where the areas between closely spaced lines are filled in with the oxide during the reflow anneal, but larger spaces or areas outside are not.

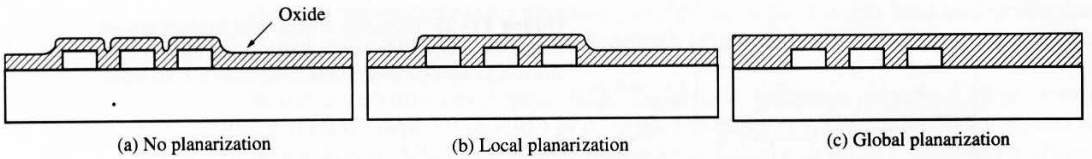

Figure 11–24 Schematic illustration of different types of planarization.

Example In Figure 11–24(b), approximately what degree of planarization has been achieved locally and what degree of planarization has been achieved globally?

Answer By Eq. (11.13), the degree of planarization is

$$\text{DOP} = 1 - \frac{x^f_{\text{step}}}{x^i_{\text{step}}}$$

The initial step height is the height of the metal line and is some finite number. In regard to local planarization, the final step height corresponds to the step height in the region that was locally planarized, that is, the regions between or within the metal lines. The step height here is close to zero, so that the degree of planarization locally is close to 1. Globally, the final step height would correspond to the maximum step height on the surface. Here it would be close to the initial step height. Therefore, the degree of planarization globally is close to zero.

Another planarization method involves etchback. In this technique the dielectric is deposited, sometimes followed by the deposition of a sacrificial layer. The overall structure is then etched back to planarize it. The earliest such process used photoresist as a sacrificial etchback layer on top of oxide. The photoresist was deposited, or spun, over the oxide layer, filling up the spaces and ending up with a nearly flat top surface. After a hard-bake, an etch is done which etches both the oxide and photoresist, ideally at the same rate. When the bottom of the photoresist is reached, a nearly flat oxide is left behind. This is illustrated in Figure 11–25, and results in local planarization, but only limited global planarization. This can also be done without a sacrificial layer if the dielectric layer is deposited thick enough. This is because the thicker a material is, the smoother the local topography tends to become. Plasma etching, argon sputtering, or chemical-mechanical polishing can be used to etch back the dielectric layer, leaving a more planar topography.

An additional planarizing method is to use a Spin-On-Glass (SOG). This is similar to the photoresist etchback technique, but here the spun-on material is left behind and serves as an intermetal dielectric. SOG materials are initially liquids, containing organic siloxanes or inorganic silicates in an alcohol-based solvent. The liquid SOG is spun onto

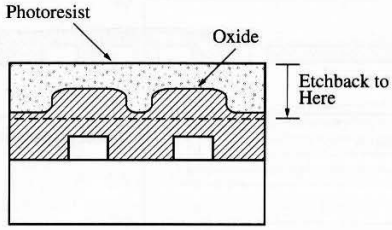

Figure 11–25 Illustration of photoresist etchback process. Photoresist is deposited over rough topography, then the structure is etched back, leaving a smooth top surface on the oxide.

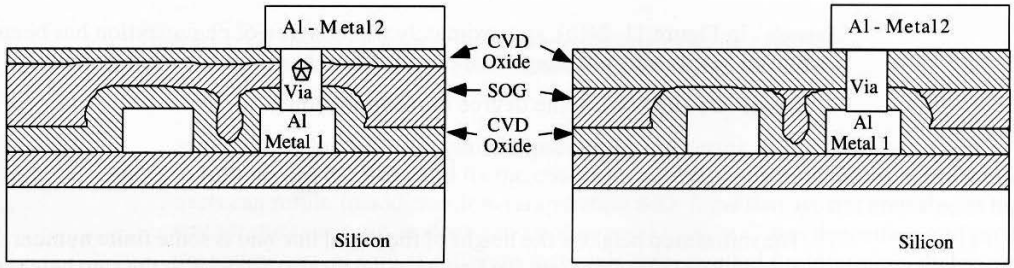

Figure 11–26 Back-end structure (left) without etchback step, showing poisoned via (star-shaped structure) where the via metal is in direct contact with Spin-On Glass (SOG) layer. The back-end structure on the right with an etchback step, does not have poisoned vias since the via metal is not in direct contact with the SOG layer.

the wafer like photoresist, filling the spaces between features. (It can also fill voids between lines as long as the voids are not completely closed off at the top.) Then the wafer is baked and cured, driving off the solvents and polymerizing the silicon-containing groups, resulting in the formation of Si-O-Si bonds. SOG is often sandwiched between layers of CVD SiO_2 in an intermetal dielectric structure, as shown in Figure 11–26. The SOG process is often combined with an etchback. After deposition and curing, the SOG is etchbacked so that no SOG remains over any metal region that will have a via made to it. This is so that when via holes are cut through the intermetal dielectric, they do not pass through any SOG, as also illustrated in Figure 11–26. If they did, residual gases and moisture from the SOG could contaminate the contacts or vias and cause large increases in contact or via resistance, so-called "poisoned vias." SOG films made from organic-based solutions, containing carbon, are more susceptible to this than those made from inorganic solutions. With SOG, local planarization and limited global planarization are achieved.

As alternatives to spin-on SiO_2 films, other Spin-On Dielectric materials (SODs) are being developed. These are either inorganic, organic polymer materials, or hybrid organic/inorganic materials (Si-O-C polymers). Table 11–6 lists some common materials. While providing good filling and local and some global planarization like SOG, they have the advantage of having lower dielectric constants than SiO_2. This means the intermetal capacitance is lower, and the *RC* delay of the interconnects is improved. The main problems with these low-*K* materials are thermal stability (i.e., they often decompose at

moderate temperatures) and moisture absorption, both of which tend to degrade their electrical properties and adhesion. Sandwich layers with CVD SiO_2 are often used to mitigate these problems.

A more recently developed oxide deposition technique is called High-Density Plasma (HDP) deposition, and was described in Chapter 9. This high-density plasma results in a dense CVD oxide film at temperatures as low as room temperature. When a separate RF bias is applied on the substrate, the angular dependence of the sputtering yield of the ions is exploited, as in bias-sputtered deposition, resulting in more planar and void-free films. HDP oxides can be doped with phosphorus and boron when used as a first level dielectric, although boron is not usually used since reasonably planar and conformal films are obtained without the need of a high-temperature reflow step. Fluorinated oxides for low-dielectric constant material can also be deposited by HDP. While HDP deposition can planarize any topography if enough film is deposited, it is usually used in conjunction with other planarizing techniques and is mainly utilized for its excellent hole or gap filling ability.

A planarization technique that has recently become part of mainstream processing is Chemical-Mechanical Polishing, or CMP. With this technique, global or near-global planarization can be achieved. As its name implies, CMP is a combination of chemical and mechanical polishing or etching. Wafers are held face down against a spinning polishing pad (made of a roughened polyurethane plate), and a colloidal silica slurry is kept between the pad and the wafer. The chemical composition and pH of the slurry are important and depend on the material being polished. For planarizing an oxide layer, a high-pH alkali-based solution is often used, while a low-pH, oxidizer-based solution is commonly used for metals. In general the chemical reaction at the wafer surface makes the surface material susceptible to mechanical abrasion by the silica in the slurry. For silicon dioxide removal, it is believed that the potassium hydroxide in the slurry reacts with the oxide to form a hydrated silicate layer, which is then removed mechanically by abrasion. For metal removal, the solution oxidizes the surface which is then polished in a similar fashion to the oxide removal.

Figure 11–27 shows a cross section of the CMP pad in contact with the top of a wafer. A semirigid polishing pad is used to allow for any bowing or thickness nonuniformities

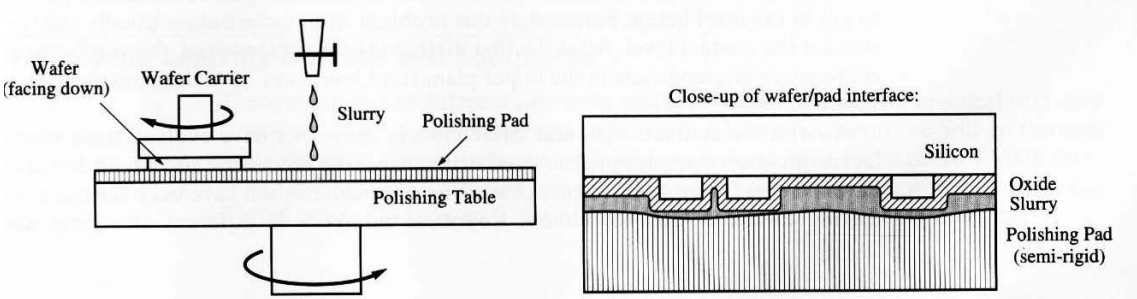

Figure 11–27 Illustration of Chemical-Mechanical Polishing (CMP) system. The right hand figure shows a close-up of wafer/pad interface.

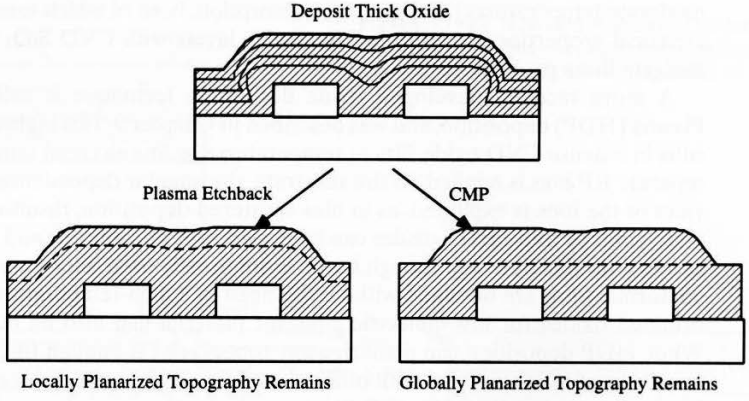

Figure 11–28 Illustrating the differences between using plasma etchback and chemical-mechanical polishing to planarize the dielectric.

in the silicon substrate. These would cause uneven etching if a perfectly rigid pad were used. The CMP process removes elevated features without much thinning of lower areas, thus planarizing the topography. CMP is used commonly for dielectric planarization, and is also being used for the W damascene process for the etchback of W (recall Figure 11–19). While standard plasma etchback processes achieve local planarization, global planarization is obtainable with CMP. Figure 11–28 illustrates this.

It should be pointed out that with all this effort at planarization, there is a drawback, other than just complicating the process. Without planarization, each layer is approximately uniform in thickness. The topography may go up and down, due to the presence or absence locally of features below it, but the vertical thickness of the layer is fairly constant (except just at the edges of features, of course). Therefore a process like etching holes in the layer is relatively straightforward since the depth one has to etch is constant from one area to the next. However, with planarization, local thicknesses of layers can vary significantly. As a result, one must etch different amounts in different locations to get to the level below. Fortunately this problem with planarization usually only occurs for the contact level. After the first dielectric level is planarized, the vias between consecutive interconnects in the upper planarized levels will have the same heights, as shown in Figure 11–1.

As with the contacts, vias, and interconnects, dielectrics have evolved from simple layers into more complex multilayered structures as the processing and circuit demands have increased. Low-temperature processing and planarization have been the two main driving forces in their evolution. Countless numbers of different structures and

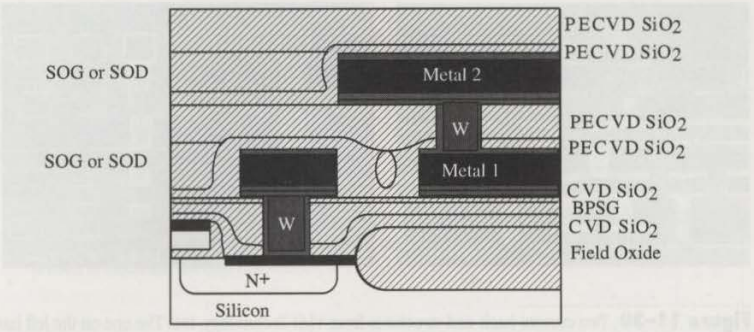

Figure 11–29 Schematic diagram of a back-end structure showing one possible dielectric multistructure scheme. Planarized topography is achieved. Other variations include HDP oxide or the use of CMP, either of which could reduce the number of sublayers —but the general structure would remain the same.

planarization schemes have been developed. Shown in Figure 11–29 is one possible, rather generalized, structure. It utilizes BPSG for the first-level dielectric, a W plug etchback, a Spin-On-Glass (SOG) or Spin-On-Dielectric (SOD) for the intermetal dielectric, and multilayer dielectric structures. The oxide layers could be either silane or TEOS based, using PECVD, bias sputtering, or HDP, and with or without an etchback or CMP. If an HDP oxide is used instead of BPSG or SOG/SOD, a moisture barrier below or above it may not be necessary. The intermetal dielectrics must be deposited and processed at temperatures less than 450°C, while the first-level dielectrics can often be done at temperatures up to 900°C or so.

Figure 11–30 shows SEM images of two current back-end structures. The first with three metal levels and 0.35-μm technology uses SOG and CMP, while the second one, a more advanced technology with five metal levels and 0.25-μm technology, uses HDP oxide and CMP.

11.3 Manufacturing Methods and Equipment

There are countless different structures and fabrication strategies for practical back-end structures, and we cannot hope to cover them all here. However, we will go through some representative processing steps and procedures used to fabricate CMOS integrated circuits. This discussion will provide some of the details that were left out in the CMOS process in Chapter 2. We will refer to that process throughout this section.