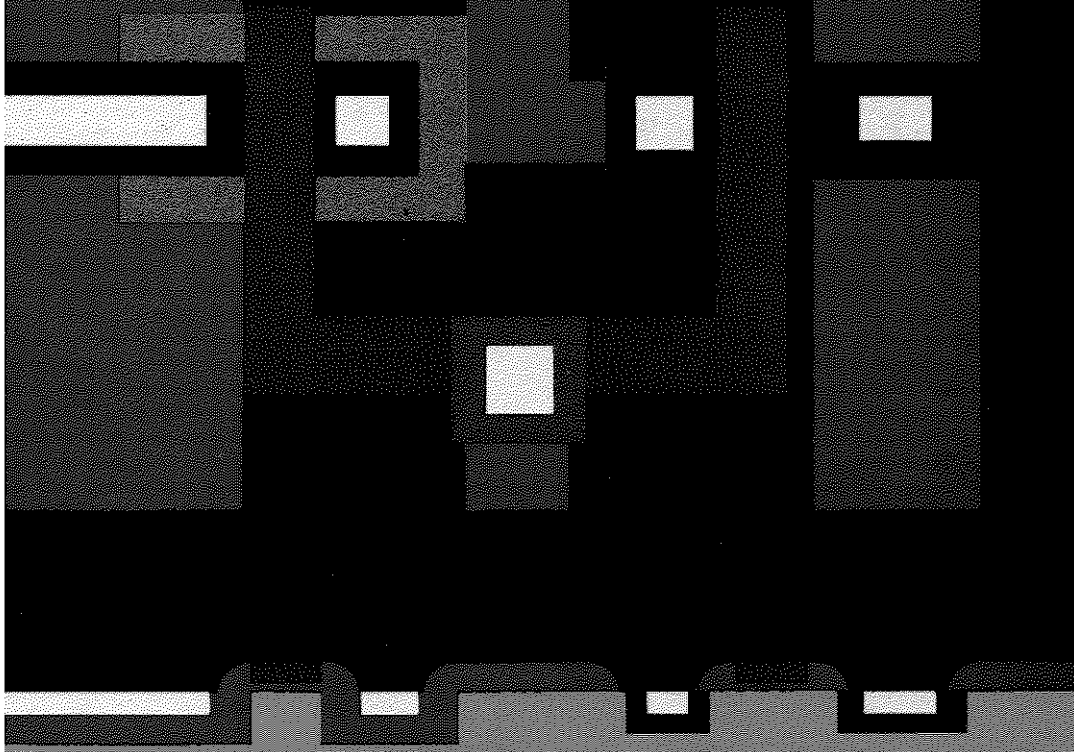
Silicon Processing for the VLSI Era

Volume 2 - Process Integration



TSMC EX. 1055 Page 1 of 7

SILICON PROCESSING

FOR

THE VLSI ERA

VOLUME 2:

PROCESS INTEGRATION

STANLEY WOLF Ph.D.
Professor, Department of Electrical Engineering
California State University, Long Beach
Long Beach, California

LATTICE PRESS

Sunset Beach, California

DISCLAIMER

This publication is based on sources and information believed to be reliable, but the authors and Lattice Press disclaim any warranty or liability based on or relating to the contents of this publication.

Published by:

Lattice Press,
Post Office Box 340
Sunset Beach, California 90742, U.S.A.

Cover design by Roy Montibon, Visionary Art Resources, Inc., Santa Ana, CA.

Copyright © 1990 by Lattice Press.
All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system without written permission from the publisher, except for the inclusion of brief quotations in a review.

Library of Congress Cataloging in Publication Data
Wolf, Stanley

Silicon Processing for the VLSI Era
Volume 2 : Process Integration

Includes Index
1. Integrated circuits-Very large scale integration. 2. Silicon. I. Title

ISBN 0-961672-4-5

9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

6.1.4.4 Cost Analysis.

• When CMOS and NMOS ICs were first being manufactured, CMOS required almost twice as many masking steps as NMOS. As NMOS processing grew more complex, however, the addition of depletion-mode loads, buried contacts, punchthrough-prevention implants, and lightly doped drains significantly increased processing costs. The costs of fabricating CMOS circuits, however, did not increase proportionately. Hence, the cost differential between the two technologies has steadily been reduced; at present the cost of manufacturing CMOS may be only 20 percent higher than that of manufacturing advanced NMOS circuits. The slightly increased cost of CMOS manufacturing is more than offset by the savings in design, packaging, system heat management, and reliability.

• The complexity of the design task is reduced in CMOS, lowering costs and allowing designs to be created more rapidly (which in turn, decreases costs further). Furthermore, since the time it takes to bring a product to market often has significant impact on market share and profitability, a shortened design time may represent a large increase in profit.

• Packaging costs can represent 25-75% of the total chip-manufacturing cost. Since the reduced power dissipation of CMOS allows the use of cheaper IC package technology, there is a significant cost savings with CMOS packaging compared to NMOS. In addition, the elimination of cooling measures and the reduced failure rates of CMOS result in lower costs. These savings translate into larger profit margins for both chip manufacturers and suppliers of electronic systems that use CMOS.

• Because grounding of the substrate can be done on the front side of the wafer, CMOS may not need back grinding or gold on the backside of the wafer, leading to further savings.

6.1.5 Disadvantages of CMOS

As is to be expected, CMOS also possesses disadvantages. Some of those listed here are inherent to all MOS circuits; these have been considered in in more detail in earlier sections on NMOS technology. Other disadvantages, however, are unique to CMOS and will be given more in-depth treatment in this chapter:

• Like NMOS, CMOS is susceptible to short-channel and hot-carrier effects when device channel lengths drop below about 2 μm (although as mentioned, the hot-carrier problem is reduced to some degree in CMOS). In addition, hot-electron effects in p-channel devices apparently do not become severe until channel lengths of below 1 μm are reached.

• CMOS has a somewhat lower packing density than NMOS.

• Static CMOS logic gates exhibit larger input capacitance than NMOS logic gates due to the additional input capacitance of the p-channel transistors, which are in parallel with the n-channel transistors.

• The need to simultaneously manufacture high-quality PMOS and NMOS devices on the same chip can give rise to processing difficulties.

• There are constraints on the scaling of PMOS devices manufactured with n^+ polysilicon gates.

• Well contacts must be provided, which takes up more chip area than required in NMOS.

• The well drive-in step requires a long process time (e.g., four hours or more at 1100°C).

• CMOS is susceptible to latchup, and hence needs guard bands or epi (see section 6.4.8). In addition, it is often very difficult to identify the exact location of the latchup site (i.e., special liquid-crystal or infra-red techniques must be used).

• Most current CMOS technologies use n^+-doped polysilicon as the gate material. An interconnect routing problem arises because the metal layer must be used when contact is made between this n^+ polysilicon and the p^+ source/drain region of PMOS devices.

• Like all other MOS technologies, CMOS is vulnerable to electrostatic-discharge damage.

6.2 THE WELL CONTROVERSY IN CMOS

There are many trade-offs involved in the optimization of a CMOS process. The choices revolve around the highly interrelated parameters of circuit performance, layout density, fabrication cost, and tolerance to latchup. As described in chapter 5, obtaining the best circuit performance from an MOS device involves maximizing the drive current and minimizing junction capacitances and body effect – all of which favor lower doping concentrations in the device body. Optimizing density, however, favors raising these same doping concentrations (to avoid punchthrough and to achieve high field thresholds). Higher density is thus achieved by allowing closer packing of adjacent n- and p-channel transistors. (Issues relating to CMOS isolation will be described in more detail in a later section.) Latchup tolerance can also be improved by spacing n- and p-channel transistors farther apart (see section 6.4), but this in turn lowers circuit density.

These complex interacting tradeoffs converge on several processing configurations that are determined at the outset of the processing sequence through the selection of the type of *well doping*. The ramifications of this choice must therefore be considered in more detail.[1,4,6]

6.2.1 The Need for Wells in CMOS

Both n- and p-channel transistors must be fabricated on the same wafer in CMOS technologies. Obviously, only one type of device can be fabricated on a given starting

TSMC EX. 1055 Page 4 of 7

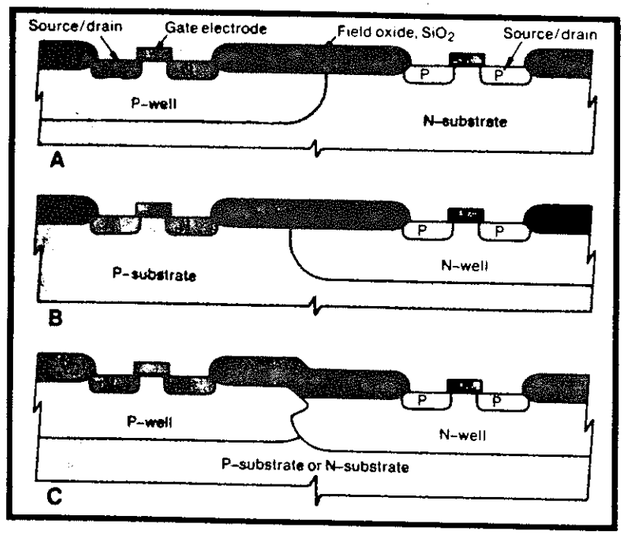


Fig. 6-8 CMOS circuits can be designed using (a) *p*-well, (b) *n*-well, or (c) twin-well technology.[3] (© 1983 IEEE).

substrate. To accommodate the device type that cannot be built on this substrate, regions of a doping type opposite that present in the starting material must be formed, as shown in Fig. 6-8. These regions of opposite doping, called *wells* (or sometimes *tubs*) are the first features to be defined on a starting wafer. This is done by implanting and diffusing an appropriate dopant to attain the proper well depth and doping profile. The well's doping type becomes the identifying characteristic of the CMOS technology.

Normally, a *p*-type substrate is connected to the most *negative* circuit voltage, and an *n*-type substrate to the most *positive*, to ensure that *pn* junctions are not forward-biased during circuit operation. Similarly, the respective well regions must also be connected to the appropriate circuit voltages to prevent forward-biasing of the junctions within the well (and between the well and substrate). Because the wells are totally junction isolated from the rest of the wafer, it is especially important that such *well contacts* be made. That is, it may still be possible to contact the substrate from the backside of the wafer even if no provision is made for a *substrate* contact from the top surface of the wafer. Such a backside connection, however, cannot be established to the *well* regions.

Because one of the device types must be located in the well, there has been some controversy as to which type of well should be used in CMOS-circuit fabrication. The performance of devices in the well will suffer as a consequence of the higher doping,

exhibiting higher junction capacitance, increased sensitivity to body effect, and decreased transconductance (due to reduced carrier mobility). Furthermore, substrate currents (e.g., due to hot-carrier effects) will be harder to collect from the well regions. The issue thus becomes that of deciding which device type should be subjected to such performance degradation.

Some argue that the NMOS devices should be built in the substrate, where their better performance can be optimized (*n*-well CMOS). This argument is persuasive for companies that have had long experience in producing high performance NMOS, since they can merely transfer this technology to the building of NMOS devices in the *p*-substrate starting material. Furthermore, if a circuit-design technique rich in NMOS devices is used (e.g., Domino logic),[97] most of the devices on the chip will be NMOS, which will again favor the building of NMOS devices in the *p*-substrate (and PMOS devices in the *n*-well).

On the other hand, the hot-electron-induced substrate current is much higher in NMOS than in PMOS and, as noted, is harder to collect from the well regions than from the substrate. In addition, device technologists might argue that the better-performing NMOS devices can afford to have their circuit behavior somewhat degraded, as this will balance the performance between these and PMOS devices.

The next sections will outline the pros and cons of both *p*-well and *n*-well configurations, as well as those of the more complex well configurations that have been developed (e.g., twin-well and retrograde-well CMOS).

6.2.2 *p*-Well CMOS

The *p-well process*, illustrated in Fig. 6-8a, involves the creation of *p*-regions in an *n*-type substrate for the fabrication of NMOS devices. The *p*-wells are formed by implanting a *p*-type dopant into an *n*-substrate, at a high enough concentration to over-compensate for the substrate doping and to give adequate control over the *p*-type doping in the well. The starting *n*-type substrate, however, must also have sufficient doping to ensure that the characteristics of devices fabricated in the substrate regions are adequate (a minimum doping concentration of of $3x10^{14}$-$1x10^{15}/cm^3$ is required). The *p*-well doping must therefore be about five to ten times higher than the doping in the *n*-substrate. If the *p*-well is doped *too* highly, however, the performance of the *n*-channel devices will be degraded through lower carrier mobility, increased source/drain to *p*-well capacitance, and increased sensitivity to body-biasing effects.

As noted in section 6.1.2, *p*-well CMOS was the first type of CMOS that could be practically manufactured. The first companies to commercially offer CMOS components produced many designs in *p*-well technology. As this experience was spread throughout the industry, *p*-well CMOS became widely established.

There are several advantages of *p*-well over *n*-well CMOS. First, *p*-well technology may be the better choice for pure-static logic, in which a good balance between the performance of both MOS device types is beneficial. Second, it is attractive for applications that require an isolated *p*-region (such as those using an *npn* bipolar transistor as an on-chip driver or *n*-channel FETs for an analog input). Third, it is less

TSMC EX. 1055 Page 5 of 7

susceptible to field-inversion problems than *n*-well CMOS, and can thus be slightly easier to fabricate. (We will describe later how *p*-well CMOS can use the well itself as a channel stop, whereas *n*-well CMOS must use a separate channel-stop process.) Fourth, if the so-called *retrograde-well* process (rather than a diffusion of a shallow implant) is used to form the wells, *p*-well technology is more feasible. It is easier to form a *p*-retrograde than an *n*-retrograde well, since boron ions penetrate deeper than arsenic or phosphorus ions at a given implant energy.

Finally, *p*-well CMOS may be better for fabricating SRAMs. Since the alpha-particle–induced soft-error rate (SER) becomes significant even in SRAMs if feature sizes are scaled to submicron dimensions, the cells should be made inside a well. The sensing of the state of an SRAM cell depends on the current provided by the cell. As a result, high-gain NMOS devices are more desirable for the pass gates and drivers in the cell, and must be built in a *p*-well.

The organization that provides university communities with IC fabrication services, the MOS Implementation Service (MOSIS), offers a standard *p*-well CMOS process (as well as standard NMOS and advanced CMOS [twin-well] processes). MOSIS cooperates with various IC-manufacturing vendors that fabricate the designs submitted to it. Circuits designed to the specifications of the standard *p*-well process can be executed by vendors, which serve as *silicon foundries* for MOSIS.

6.2.3 *n*-Well CMOS

In the *n*-well process, shown in Fig. 6-8b, the *p*-channel devices are formed in the more heavily doped *n*-well. As noted earlier, *n*-well technology became the choice of companies with extensive experience in building NMOS ICs.[8] Because the NMOS device could be fabricated in a lightly doped substrate, virtually all of the experience that had been amassed in fabricating high-performance NMOS could be transferred to an *n*-well CMOS process. As a result, virtually all EPROMs, microprocessors, and dynamic RAM designs in the generations of technology built with 1.25-2.0 μm dimensions were implemented with *n*-well CMOS.

This technology does have some disadvantages. First, as mentioned earlier, it is more sensitive to field-inversion problems than *p*-well CMOS. Second, it may be more difficult to build pure-static, high-performance logic circuits with *n*-well CMOS.

EXAMPLE 6-1: An *n*-well CMOS process is to be developed for operation with a power-supply voltage of V_{DD} = 5 V. The substrate doping of the *p*-type wafers is $1x10^{15}/cm^3$. The *n*-wells are to have an average dopant concentration of $1x10^{16}/cm^3$. The *p*-channel MOSFET sources and drains are to have junction depths of 0.4 μm and an average dopant density of $10^{18}/cm^3$. What is the minimum *n*-well depth needed to avoid vertical punchthrough to the substrate?

SOLUTION: Vertical punchthrough will occur if the depletion region of the source/drain-to-well junction were to contact the depletion region of the well-substrate junction when V_{DD} = 5 V (see section 5.5.2).

The source-to-well junction is essentially a one-sided *pn* junction with a built-in voltage of V_{bi} = 0. 82 V. From Fig. 5-20, we estimate that the depletion-region width extends into the *n*-well ~0.35 μm, since there is no applied voltage across the junction (i.e., both the source and the well are connected to V_{DD}). The *np*-junction to the substrate has a built-in voltage of 0.63 V, and the total depletion width of this junction at a 5 V bias is ~1.9 μm. We calculate that about 0.19 μm of this depletion region is in the *n*-well. The *n*-well must therefore be deep enough to accommodate the depth of the source junction (0.4 μm) as well as the sum of the depletion region widths in the well in order for vertical punchthrough to be avoided. While the total of these dimensions is 0.94 μm, it is good engineering practice to increase the depth of the well by about 50% to account for process variations. A reasonable well depth might therefore be 1.5 μm.

6.2.4 CMOS on Epitaxial Substrates

As will be described in the section dealing with latchup prevention (section 6.4.8.2), heavily doped substrates with a more lightly doped surface epitaxial layer have been utilized to suppress latchup in CMOS.[11] When such starting material is used with single-well CMOS, the epitaxial layer is doped to a concentration equal to that of the substrate in a nonepitaxial wafer used for that process. If a twin-well CMOS process is used, the epitaxial layer is doped to a level significantly lower than that required for building either the *p*- or *n*-channel MOSFETs (see section 6.2.5).

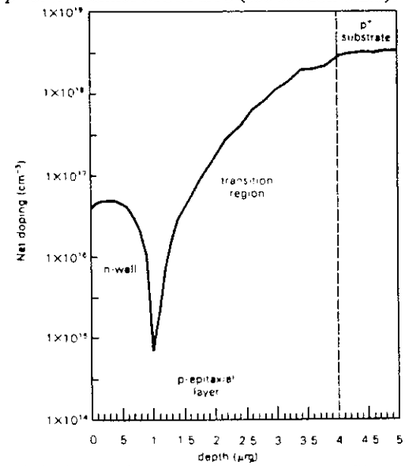


Fig. 6-9 A *p*-on-*p*-epi doping profile and diffused *n*-well profile measured after the entire CMOS process has been completed. The as grown epi thickness is 4 μm and the final lightly doped epi thickness is about 2 μm.[116] (© 1987 IEEE).

The epitaxial layer is made thicker than the well depth, since the dopants in the heavily doped substrate under the epi layer diffuse toward the surface as the well dopants are diffused toward the bulk (Fig. 6-9). Thus, some of the epitaxial layer becomes more heavily doped during the CMOS process flow. The process is designed so that the bottom of the well is eventually adjacent to the heavily doped substrate region.

Either n-epi on n^+ substrates or p-epi on p^+ substrates can be used, with each method having advantages and drawbacks. Because the problems with n-epi on n^+ are more serious, p-epi on p^+ is more widely used. The major limitation of the latter approach is that the outdiffusion of boron from the p^+ is much more severe than it is in n-epi on n^+. (The reason for this is that boron diffuses much more rapidly than antimony, which is the most widely-used n^+ dopant material.) Thus, a thicker p^-- epitaxial layer must be used.

In addition, the transition region between the p^+ substrate and the p^-- epi layer is thicker, producing a larger series resistance (R_{sub}), which in turn reduces latchup immunity. On the other hand, the p-on-p^+ material is less sensitive to process-induced defects, and the p-type substrate provides higher conductivity under NMOS devices. Such additional conductivity is desirable, since it reduces the voltage drops caused by the substrate currents (generated as a result of the hot-carrier effects in short-channel devices). It is especially important in the regions containing NMOS devices, since the hot-electron substrate current is much higher in such devices than in PMOS devices.

The n-on-n^+ epitaxial substrates also offer some advantages. First, SRAMs are often built on n-type substrates, because the p-well to n-substrate junction provides protection from radiation-induced discharging of the memory's n-type storage nodes in the p-well.[12] Second, retrograde p-wells are easier to implement because their implantation energy requirements are much lower. Third, the n-to-n^+ transition region is smaller than that in p-on-p^+ substrates, and the smaller value of R_{sub} provides improved latchup protection.

The limitations of n-epi on n^+ involve the process by which the heavily doped substrate is grown. Antimony (Sb) is the n-type dopant used, both because its diffusion coefficient is so low and because it exhibits much less lateral autodoping than arsenic (the other slow-diffusing n-type dopant). The problem with Sb is that its segregation coefficient, k_o, is very small (i.e., $k_{o,Sb} = 0.023$; see, Vol. 1, chap. 1), and thus a large quantity of Sb must therefore be put into the Si melt to ensure that a sufficiently heavily doped ingot will be produced. In even the most highly refined Sb there are high concentrations of unwanted heavy metals, which become incorporated into the growing silicon crystal. In addition, the oxygen content of the Sb-doped crystal is relatively small, due in part to the special growing conditions used when the Sb-doped ingot is pulled.[7,9,10] As a result, intrinsic gettering techniques that would getter the metals in the substrate are not as effective as they are in p-on-p^+ epi. For these reasons, n-on-n^+ substrates are less frequently chosen when epi-CMOS is implemented.[12]

A problem that exists with both types of epitaxial substrates for CMOS is that the wells cannot be made too deep, since the lateral diffusion would then take up too much area. On the other hand, if the wells are too shallow, vertical punchthrough will ensue.[13] A second problem is that of leakage current. Appreciable leakage current can

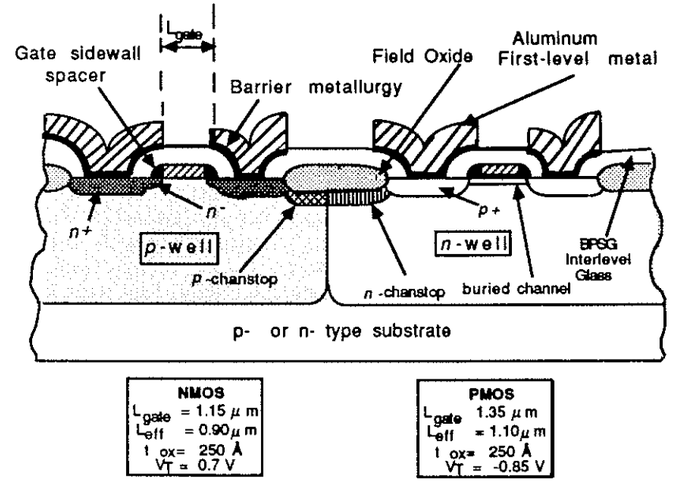


Fig. 6-10 Cross section of a twin-well 1.2-μm CMOS structure.[12] Reprinted with permission of Semiconductor International.

flow in the vertical path in two different situations. First, punchthrough can occur if the depletion regions of the p^+/n-well junction and the n-well/p^+-substrate junction touch each other. The problem is even more severe in the case of the heavily doped p^+ substrate, due to the high degree of boron outdiffusion. A major limitation is thus imposed on the minimum epitaxial-layer thickness. Second, it may be possible for the source regions to become biased to a potential below V_{SS}.

Another problem of epi-CMOS is back surface autodoping. For example p-type 20 W-cm ($7x10^{14}$ boron/cm^3) epi on a 0.005 Ω-cm ($2x10^{19}$ boron/cm^3) substrate is representative of epi for CMOS devices. If the substrate back is not sealed (e.g., using a sealing layer such as undoped silicon dioxide or silicon nitride), boron evaporation can contribute to autodoping on the front surface during the entire epi deposition cycle. This widens the epi/substrate interface and may even prevent the epi from reaching the 20 Ω-cm specified resistivity. Note that if silicon nitride is used as a sealing layer it should only be used in thin layers (e.g., less than 100 nm thick) since its high intrinsic stress causes it to bow the silicon.[113]

6.2.5 Twin-Well CMOS

With the twin-tub approach, two separate wells are formed for n- and p-channel transistors in a lightly doped substrate (Figs. 6-8c and 6-10). The substrate may be either a lightly doped wafer of n or p material, or a thin, lightly doped epitaxial layer on

TSMC EX. 1055 Page 7 of 7