

5.6.4 Techniques for Reducing Hot-Carrier Degradation

Device performance degradation from hot electron effects can be reduced by a number of techniques. These include the following:

- The voltages applied to the device can be decreased (e.g., by lowering the power-supply voltage from 5 V to 3.3 V). The decision to implement this reduction, however, is not in the hands of the device designer or fabricator. The issue of reduced supply voltages for submicron MOS technologies is covered in chapter 6, section 6.7.2.
- The time the device is under the voltage stress can be shortened (e.g., by using a lower duty cycle and clocked logic).
- Appropriate drain engineering design techniques, which results in special drain structures that reduce hot electron effects in MOS devices (i.e., double-diffused drains and lightly-doped drains), can be implemented.
- The density of trapping sites in the gate oxide can be reduced through the use of special processing techniques.
- Thin, lightly doped epitaxial-layers on low-resistance substrates can be employed to shunt away substrate current and help eliminate the problem of impact-ionization-induced latchup.

We will next discuss the details of the latter three approaches since their implementation is the task of device and process engineers.

5.6.5 Lightly Doped Drains (LDD)

It has been determined that hot-carrier effects will cause unacceptable performance degradation in NMOS devices built with conventional drain structures if their channel lengths are less than $2\text{ }\mu\text{m}$. To overcome this problem, such alternative drain structures as *double-diffused drains* and *lightly doped drains* (LDDs) must be used. The purpose of both types of structures is the same – namely, to absorb some of the potential into the drain and thus reduce E_M . Double-diffused drains, however, are less effective for short-channel devices (i.e., $\leq 1.25\text{ }\mu\text{m}$) because they cause deeper junctions and more overlap capacitance. We will thus restrict our discussion to LDDs. Table 5-2 shows the evolution in AT&T's Twin-Tub CMOS technology, with respect to such drain structures.³⁰

In the LDD structure, the drain is formed by two implants (Fig. 5-30). One of these is self-aligned to the gate electrode, and the other is self-aligned to the gate electrode on which two oxide *sidewall spacers* have been formed. (An extensive report on the details of sidewall spacer technology for both MOS and bipolar devices can be found in ref. 66.) The purpose of the lighter first dose is to form a lightly doped section of the

Table 5.2 Evolution of Device Structures in AT&T's Twin-Tub CMOS Technology Development³⁰ (Twin-Tub VI announced in 1989)⁸³

	Twin-Tub I	Twin-Tub II	Twin-Tub III	Twin-Tub IV	Twin-Tub V	Twin-Tub VI
Design Rule	3.5 μm	2.5 μm	1.75 μm	1.25 μm	0.9 μm	0.6 μm
L_{eff}	2 μm	1.5 μm	1.3 μm	1.0 μm	0.75 μm	0.4 μm
t_{ox}	600 Å	350 Å	250 Å	200 Å	150 Å	125 Å
Device Structure	Conventional	Conventional	DDD	LDD	N&P LDD	N&P LDD

drain at the edge near the channel. In NMOS devices, this dose is normally $1\text{--}2 \times 10^{13}$ atoms/cm² of phosphorus.

The value of E_M is reduced by this structure because the voltage drop is shared by the drain *and* the channel (in contrast to a conventional drain structure, in which almost

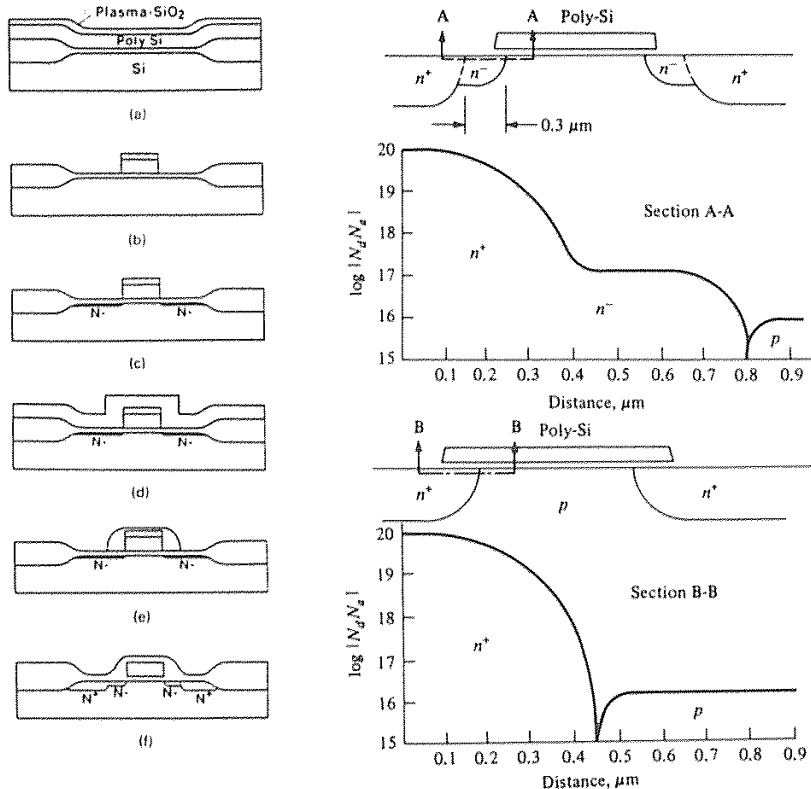


Fig. 5-30 (a) Process sequence used to form lightly doped drain (LDD) structures. (b) Doping profile in an LDD structure taken through section A – A. (c) Doping profile in a conventional drain structure taken through section B – B.

the entire voltage drop occurs across the lightly doped *channel* region). Figure 5-25 shows the electric-field profile at the drain of a MOSFET, both with and without an LDD structure.³⁸ We see that the electric field can be reduced by about 30-40%. Since the hot-electron-induced gate currents are exponentially dependent on E_M , this is sufficient to reduce currents by many orders of magnitude. As a result, the stability of the device is greatly increased. To ensure a high quality interface under the sidewall-spacer oxide, it is important that the thermally grown gate oxide remain in place after the polysilicon gate etch. This requires a polysilicon etch process with high selectivity to oxide, since the gate oxide is typically less than 20 nm.

The heavier, second dose forms a low resistivity region of the drain region, which is also merged with the lightly doped region. (In NMOS devices this implant is typically arsenic at a dose of about 1×10^{15} atoms/cm².) Since it is further away from the channel than would be the case in a conventional drain structure, the depth of the heavily-doped region of the drain can be made somewhat greater without adversely impacting the device operation (e.g., 0.3 μm deep versus 0.18 μm deep). The increased junction depth lowers both the sheet resistance and the contact resistance of the drain (see chap. 3). Deeper junctions also provide better protection against junction spiking.

The disadvantages of LDD structures are their increased fabrication complexity compared to conventional drain structures and the increased parasitic resistance of the source and drain regions caused by the lightly doped regions of the drain. This increase in parasitic channel resistance results in devices that dissipate more power for a constant applied voltage.

The effect on the I_D - V_{DS} curves of the MOSFET due to the additional voltage drop across the two lightly doped regions is that I_D does not saturate until a higher value of V_{DS} is applied. Due to the extra series resistance, the total channel conductance is appreciably lower in the linear region, but only slightly lower in the saturation region. This is because the channel region in saturation already has a high channel resistance, while in the linear region the resistance is much lower. The additional series resistance of the LDD therefore does not significantly increase the total resistance of the MOSFET in saturation, and I_D remains less affected. Consequently, the drain current is reduced more significantly in the linear region than in the saturation region.)

The I-V curves for LDD devices are consequently recognizable by their characteristically round shape, as shown in Fig. 5-31, curve 6. These curves also illustrate progressively the other short-channel effects described in section 5.5, beginning with the I-V curve of an ideal long-channel MOS device (Curve 1). Thus, curve 6 shows the I-V curve of a short-channel MOSFET with an LDD.

5.6.5.1 Drain Engineering for Optimum LDD Structures. A proper LDD design should provide adequate hot-carrier protection while not introducing excessive source/drain resistance, which would degrade the device performance. The design and fabrication effort undertaken to achieve such an optimum LDD structure is referred to as *drain engineering*; this as we shall see, is a relatively complex task. A thorough approach to designing adequate hot-electron protection would entail addressing all of the following objectives:

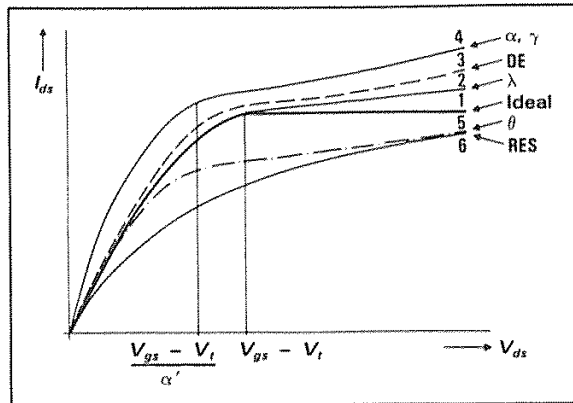


Fig. 5-31 Progressive influence of short-channel effects and LDD on the characteristics of an ideal, long channel MOSFET (Curve 1). Curve 6 shows the I-V curve of a short channel MOSFET that is also fabricated with an LDD, while Curve 5 shows the characteristic of the same device without an LDD.²⁸ (© 1986 IEEE).

1. Reducing the maximum electric field in the silicon as much as possible.
2. Ensuring that the injection position (i.e., the E_M point) is located under the gate edge.
3. Ensuring that the impact ionization region is pushed far below silicon surface to reduce the possibility of hot carriers reaching the Si-SiO₂ interface.
4. Separating the point where the electric field in the silicon is a maximum from the point of maximum current flow in the channel.
5. Minimizing the increase in the parasitic resistance due to the LDD structure.

The degree by which E_M is reduced (*Objective #1*) depends primarily on the doping value of the lightly doped extension of the drain, as well as on its length. If the doping level is too high, the value of E_M is not sufficiently reduced, and hot-electron protection is not provided. If the level is too low, the drive-current capability of the device is severely reduced, and the surface will easily be depleted by the hot carriers that do get trapped in the gate oxide above it (i.e., the device will again be vulnerable to hot-electron degradation). A model which calculates the electric field in LDD structures has been published.⁷² It concludes that the lightly doped region should be long enough to attenuate the electric field to a value that is below the critical ionization field, but should still be short enough to keep the series resistance from becoming excessive. In early LDD structures in NMOS this meant that the primary parameters selected were the n^- length and its doping concentration.³⁹

A process for forming LDD structures in *either* or *both* PMOS and NMOS structures in CMOS with only two photomasks was reported by Parrillo, et al.⁴⁰ An extension of this process also uses removable TiN gate sidewall spacers and incorporates self-alignment of the lightly doped regions to their respective gates without overcompensation (Fig. 5-32). Another example LDD structure in 0.8- μm CMOS is given in reference 41.

To keep the location of the E_M point under the gate (*Objective #2*) and yet maintain high drivability requires the proper combination of gate-drain (g/d) overlap length and spacer length. The g/d overlap length should be long enough to let the E_M point lie under the gate, and the spacer length should be short enough to let the n^+ region reach under the gate. In general, for process controllability, if a 0.3- μm -thick polysilicon film is used for the gate, a 0.2- μm -wide spacer is selected.

To attain *Objective #3*, a metal-coated LDD structure has been developed.⁴² It uses a deeper n^- phosphorus profile than the n^+ source-drain (arsenic), in order to steer the maximum current path away from E_M (Fig. 5-33b).

A buried LDD structure has also been proposed to reach *Objectives #3 and #4*. This structure uses a retrograde LDD profile, with the peak concentration below the Si surface (Fig. 5-33c). Besides a reduction in substrate current similar to the metal-coated LDD, this profile also suppresses hot-carrier injection by driving the current away from the surface and shifting the avalanche region further into the bulk silicon.

The buried LDD approach together with an additional shallow arsenic implant (Fig. 5-34d), may allow NMOS transistors of 0.6 μm to be built with adequate hot-electron protection and with minimal the increase in parasitic resistance due to the LDD structure (*Objective #5*).^{30,43,44} The idea is to form an abrupt profile at the silicon surface underneath the spacer to reduce the series resistance. The buried LDD profile is

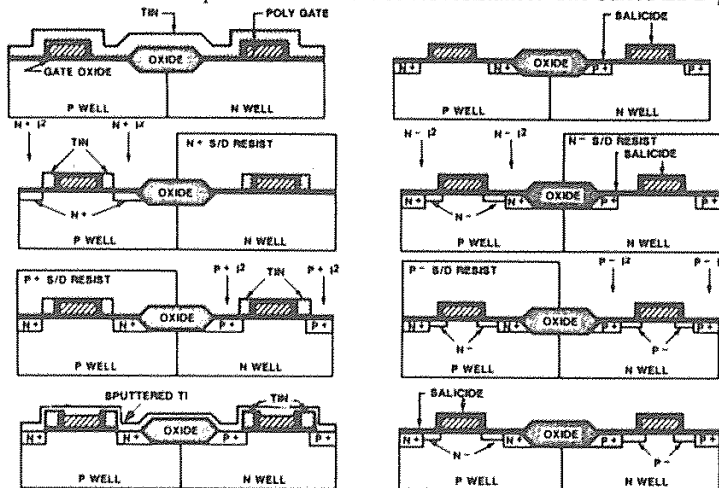


Fig. 5-32 Process sequence showing the TiN removable spacer process.⁸² (© 1989 IEEE).

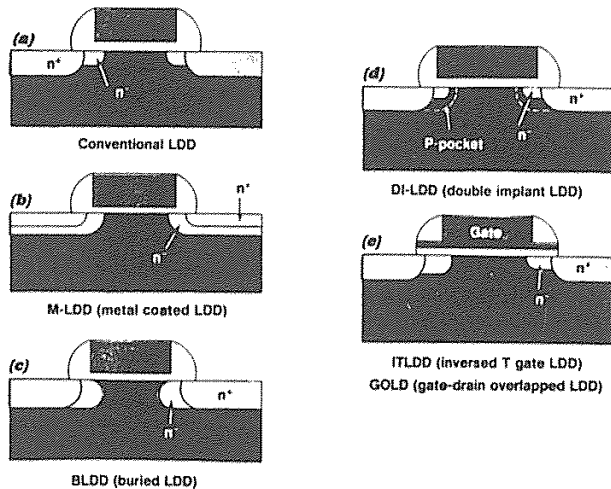


Fig. 5-33 Conventional LDD drain engineering and some of its variations — all designed to reduce hot-carrier effects — are shown.³⁰ (© 1986 IEEE).

used to force the main current path and the impact-ionization region deep into the silicon. There are two peaks in the lateral electric-field distribution, and the maximum current path is through the saddle point of these two high field regions. Thus, this device structure improves both the current drive and the hot-carrier resistance. A further modification of the buried LDD is the *sloped-junction LDD (SJLDD)*, shown in Fig. 5-33e.⁴⁵ In this structure a 165-keV phosphorus implant is used to form the lightly doped drain region; this provides improved device lifetime under high field stress. The main reason for the improvement is that hot-carrier generation is driven further away from the Si-SiO₂ interface by the junction of the SJLDD.

5.6.5.2 Asymmetrical Characteristics of LDD MOSFETs. Formation of the lightly doped regions of LDDs is accomplished by means of ion implantation, with the polysilicon gate used as the implant mask. This can cause asymmetric doping of the source and drain regions,^{46,64} with such asymmetry occurring as a result of implanting off-axis (typically, at an angle of 7°) to avoid channeling (see Vol. 1, chap. 9). This produces lateral shadowing (S in Fig. 5-34) of the substrate on one side of the poly gate, and penetration of dopants through the leading corner of the poly on the other side. The problem is compounded by etch processes that produce a re-entrant angle in the poly gate sidewall (Fig. 5-34 - poly etch process A.) The impact of this effect on the device characteristics can be examined by considering the implants to the lightly doped and the heavily doped regions separately.

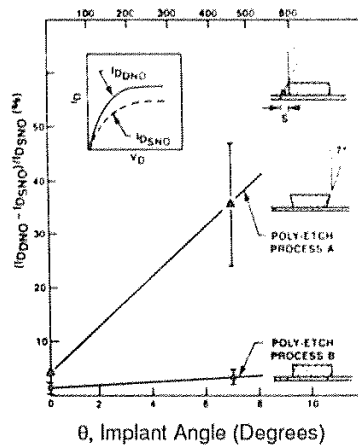


Fig. 5-34 Percent asymmetry in saturated drain currents ($V_{DS} = V_{GS} = 5$ V) versus source/drain implant angle for LDD NMOS devices having reentrant- and vertical-wall polysilicon gates. Insert schematically shows source-non-overlap (SNO) and drain-overlap (DNO) I-V behavior.⁴⁰ (© 1986 IEEE).

If the implant that forms the lightly doped region of the drain (usually phosphorus in NMOS devices) is done off-axis, the as-implanted region in the shadow of the gate will have its edge displaced from the gate edge. As a result, there will be *no overlap* of the implanted region and the gate edge. Even after a drive-in step, the overlap of this edge with the gate will be less than the overlap of the nonshadowed implant region edge and the gate. If the less-overlapped side is used as the source, the extra resistance introduced will reduce the drain current ($I_{D_{sno}}$ in Fig. 5-34), thus degrading circuit speed. However, if the drain side has less overlap, the extra resistance will not affect I_D (curve $I_{D_{dno}}$ in Fig. 5-34). The asymmetry in drive current when the same device is connected in these two ways can be as much as 40%; this can be catastrophic in circuits that require closely matched device characteristics.

The problems arising from such n^- implant-shadowing can be alleviated by one of the following measures:

- Insuring that the poly-etch process produces vertical sidewalls in the poly gates.
- Avoiding excessive reoxidation of the polysilicon after its definition (note that this also prevents a gap from being established between the drain region edge and the edge of the gap). A dry-oxidation step is often employed to reduce the accelerated oxidation rate on n^+ polysilicon.
- Implanting with a vertical implant through a screen oxide²⁷ (which reduces the channeling that would otherwise occur when a vertical implant is performed).

- Using a poly-etch process that produces a slight positive bevel to the poly sidewall profile, and combining this etch with a smaller angle (3° vs 7°) off-axis implant.⁴⁶

Let us next examine how the implantation of the heavily doped region impacts the structure of the LDD. If an off-axis implant is again used with the heavily doped implant, asymmetric implant effects can again alter the LDD structure. That is, on the side of the gate that is *not* shadowed, the arsenic atoms penetrate the leading corner of the spacer and thereby enter the substrate under the spacer to some extent. If the spacer is too narrow, this can wipe out the lightly doped drain-extension region. When this happens, the device can lose the hot-carrier protection that was to be provided by the LDD structure, and it will become vulnerable to degradation by hot-carrier effects. To avoid this, the spacer must be sufficiently wide. Some guidelines for choosing the proper spacer width to deal with this problem are given in reference 47.

The general problem involving the gate-to-drain overlap with respect to its impact on the MOSFET characteristics was reviewed by Ko, et al.⁴⁸ It was observed that the critical dimensions in drain structures having weak overlap (WO) to the gate were only tens of nm in devices in which $L_{\text{eff}} = 1 \mu\text{m}$. In general, the drain current decreases as the overlap is weakened, a double hump appears in the substrate current of asymmetrical WO devices, and the reliability of the WO devices can be lower than devices having adequate overlap. To avoid random degradation of device performance by WO effects that arise as a result of process variations, stringent process-control measures must be exercised when submicron devices are fabricated.

5.6.6 The Impact of IC Processing on Hot-Carrier Device Degradation

The lifetime of a device in which the hot-electron degradation impacts device performance can be increased by keeping the number of trapping centers in the gate oxide to a minimum. In essence, this reduces the density of states that the hot electrons injected into the oxide can occupy. Maintaining a minimum number of trapping centers can be achieved in several ways during the device-fabrication process sequence. First, the gate oxide should be grown by a process that produces low interface-trap densities (such processes are described in Vol. 1, chap. 7). It has also recently been reported that the incorporation of optimized amounts of fluorine during the gate-oxide growth process suppresses interface-state generation during injection.⁴⁹ Second, the degradation of the oxide caused by damage during process steps carried out in a plasma environment should be minimized (damaging processes include plasma-enhanced CVD of oxides and nitrides, RIE, and sputter deposition of metal). Finally, the amount of hydrogen that is incorporated at the Si-SiO₂ interface should be reduced. The latter two topics will be considered in more detail here.

In typical CMOS processing, most of the damage created by RIE or ion implantation is annealed out at a high temperature (above 800°C) before metallization. Once the first layer of aluminum is deposited, however, the maximum annealing temperature is

limited to $\sim 450^\circ\text{C}$. It has therefore been reported that the RIE of the second layer of metal in multilevel metal processes deteriorates the device-aging characteristics. It has been found that while a subsequent anneal performed at 375°C is ineffective in annealing out this damage,⁵⁰ an anneal at 450°C improves device-aging characteristics.

Excess hydrogen at the Si-SiO₂ interface is also reported to be a culprit in increasing the density of the interface states. In most cases, hydrogen is used to fill the dangling bond, forming Si-H at the Si-SiO₂ interface. However, the Si-H bond can be easily broken by injected hot electrons. Furthermore, excess hydrogen introduced during processing can diffuse to the interface and lead to enhanced bond-breaking behavior. Such excess hydrogen can arrive during the final sinter step before passivation (i.e., in H₂ or N₂ + 5% H₂), or during the deposition of a conventional silicon nitride passivation layer. It has therefore been reported that sintering in pure N₂ produces devices with a lower device degradation rate.⁵¹ Similarly, it has been reported that when a fluorinated nitride (F-SiN) is used as a passivation layer,^{50,52} devices exhibited slower degradation rates than those seen in devices with conventional SiN passivation layers. F-SiN films can be produced by incorporating NF₃ in the deposition process to form an Si-F bonding structure instead of Si-H in this film.

The presence of fluorine in the gate oxide (possibly inadvertently originating from a BF₂⁺ source/drain implant implant) has recently been reported to increase the hot electron resistance of devices fabricated with such oxides.⁷⁴ Other reports confirm that a deliberate incorporation of fluorine into the gate oxide (from implanting fluorine into a polysilicon film, and then diffusing it into the gate oxide) produces a more hot-electron resistant interface.^{78,79}

5.6.7 Hot-Carrier Effects in PMOS Transistors

Hot-carrier effects are not significant in PMOS transistors at channel lengths greater than $1\ \mu\text{m}$ because the impact-ionization rate of holes is 3 to 4 orders of magnitude lower than that of electrons at a given electric field.⁵³ At submicron channel lengths, however, hot-electron effects in PMOS start to become important. It has been reported that two hot-carrier effects predominate in such PMOS devices. Both are caused by hot electrons that are generated by impact ionization and are then injected into the oxide, becoming trapped. Hot holes do not appear to cause significant effects unless a device is stressed at large magnitudes of V_G . In PMOS devices, the gate-bias polarity favors electron injection (which is opposite to that in NMOS) but retards the injection of holes.

In the first degradation effect, these electrons are trapped near the drain and shorten the effective channel length. This reduction in L_{eff} is even more severe in the buried-channel PMOS devices (used in CMOS technologies that have an n^+ polysilicon gate), because their buried-channel nature makes them more vulnerable to source-drain punchthrough,⁵⁴ and consequently increases subthreshold leakage. This effect is called *hot-electron-induced punchthrough* (HEIP).

The second effect was observed in PMOS devices with p^+ , as well as n^+ , polysilicon gates. In these devices, the injected and trapped hot electrons also produce V_T shifts

(which reduce the magnitude of V_T) and increases in the transconductance, g_m . Nevertheless, contrary to what occurs in NMOS, in PMOS these effects tend to saturate. This is explained by the fact that the electrons trapped near the drain reduce the electric field present there in PMOS devices.⁵⁵ As a result, this effect does not appear to limit PMOS devices fabricated with p^+ polysilicon gates as long as L_{eff} is $\geq 0.6 \mu m$.⁵⁶ Still, LDD structures appear to be needed for p -channel devices fabricated with n^+ polysilicon gates when L_{eff} gets to be $0.8 \mu m$ or less.^{54,57}

The LDD structure also helps to reduce subthreshold leakage in PMOS devices, where excessive junction depths caused by lateral diffusion have been a problem. The use of LDD basically increases L_{eff} , hence compensating for the L_{eff} reduction caused by hot-electron injection. A new LDD structure for PMOS devices, called a *halo LDD*, is described in reference 63. In this structure, a deeper phosphorus implant is placed below the lightly doped drain-extension p -type implant. The punchthrough resistance of the PMOS device is reported to be significantly improved by this LDD structure.

5.6.8 Gate-Induced Drain-Leakage Current

Another type of leakage current between drain and substrate in thin gate-oxide (12-28 nm) MOS devices has been observed at drain voltages much lower than the breakdown voltage.^{58,59} This is not really a hot-electron-induced effect, but since it uses LDD structures to reduce its magnitude, we describe it in this section. The basis of this current is band-to-band tunneling that occurs through the gate oxide into the deep depletion layer in the gate-to-drain overlap region. It has been reported that in order for this leakage current to be limited to less than $0.1 \text{ pA}/\mu m^2$, the oxide field in the gate-to-drain overlap region must be limited to 1.9 MV/cm . LDD structures are effective in reducing this leakage current, but only if the doping in the n^+ extensions is less than $10^{18}/\text{cm}^3$. Buried LDD structures with the peak of the n concentration at several hundred angstroms beneath the Si-SiO₂ interface appear to be even more effective than conventional LDD structures in reducing the effect. This effect is predicted to become more of a problem as devices continue to be scaled - that is, if a 10-nm-thick gate oxide is used, the E-field will be less than 1.9 MV/cm in the oxide only if a 3.1 V power supply is used.

REFERENCES

1. R. S. Muller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 2nd Ed., New York, John Wiley & Sons, 1986.
2. R. F. Pierret, *Field-Effect Devices*, Vol. IV in the Modular Series on Solid-State Devices, Reading, MA, Addison-Wesley, 1983.
3. D. K. Schroder, *Advanced MOS Devices*, Vol. VII in the Modular Series on Solid-State Devices, Reading, MA, Addison-Wesley, 1987.
4. Y. P. Tsividis, *The Operation and Modeling of the MOS Transistor*, New York, McGraw-Hill Book Co., 1987.

5. E. H. Nicollian and J.R. Brews, *MOS Physics and Technology*, Wiley-Interscience, New York, 1982.
6. S. M. Sze, *Physics of Semiconductor Devices*, 2nd Ed. Wiley, New York, 1981.
7. D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, McGraw-Hill Book Co., New York, 1983.
8. R. A. Chapman et al., *Tech. Dig. IEDM*, 1987, p. 362.
9. R. J. Bayruns et al., *IEEE J. Solid-State Circuits*, SC-19, 1984, p. 755.
10. H. L. Armstrong, "A Theory of Voltage Breakdown of Cylindrical P-N Junctions," *IRE Trans. Electron Dev.*, ED-4, 15 (1957).
11. A. S. Grove, *Physics and Technology of Semiconductor Devices*, Wiley, New York, 1967, Chap. 6
12. E. C. Douglas, "Advanced Process Technology for VLSI Circuits," *Solid State Tech.*, 24, May 1981, p. 65.
13. C. M. Osborn and E. Bassous, *J. Electrochem. Soc.*, 122, 89, (1975).
14. S. M. Sze, *Semiconductor Devices: Physics & Tech.*, Wiley, New York, 1985.
15. N. Kotani and S. Kawazu, *Solid-State Electronics*, 22, p. 63, (Jan. 1979).
16. T. Yamaguchi et al., *IEEE Trans. Electron Dev.*, ED-31, 1984, p. 205.
17. C. Mead and L. Conway (Eds.), *Introduction to VLSI Systems*, Reading, MA, Addison-Wesley, 1980.
18. A. Mukherjee, *NMOS and CMOS VLSI Systems Design*, Prentice-Hall, Englewood Cliffs, N.J., 1986.
19. N. Kotani and S. Kawazu, *Solid-State Electronics*, 22, 63 (January 1979).
20. L. D. Yau, *Solid-State Electronics*, 17, 1059, (1974).
21. L. M. Dang, *IEEE J. Solid-State Circuits*, SC-14, 358 (1975).
22. G. Merkel, *Process and Device Modelling for IC Design* (F. Van de Wiele, W. L. Engle, and P.G. Jespers, Eds.), Noordhoff, Leyden (1977), p. 705.
23. R. R. Troutman, "VLSI limitations from Drain-Induced Barrier Lowering," *IEEE Trans. Electron Dev.*, ED-26, p.461, 1979.
24. J. Zhu et al., "Punchthrough Current for Submicron CMOS VLSI," *IEEE Trans. Electron Dev.*, Feb. 1988, p. 145.
25. K. M. Cham, et al., *Computer Aided Design and VLSI Device Development*, Kluwer Academic Publishers, Boston, 1986, p. 194.
26. H. Nihara, et al., "Anomalous Drain Current in NMOS and its Suppression by Deep Ion Implantation," *IEDM Tech. Dig.*, p. 487 (1978).
27. L. C. Parrillo, "CMOS Active and Field Device Fabrication," *Semiconductor International*, April 1988, p. 64.
28. C. Duvvury, "A Guide to Short-Channel Effects in MOSFETs," *IEEE Circuits and Systems Magazine*, Nov. 1986, p. 6.
29. M. Kakamu et al., "Power Supply Voltage for Future CMOS VLSI in Half- and Sub Micrometer," *Tech. Dig. IEDM*, 1986, p. 399.
30. M. L. Chen, "CMOS Hot-Carrier Protection with LDD," *Semiconductor Internat.*, April 1988, p. 78
31. P. K. Ko, "Hot Electron Effects in MOSFETs," Doctoral Thesis, Dept. of EECS, University of California, Berkeley, June 1972.

32. T. Y. Chan, P.K. Ho, and C. Hu, *IEEE Electron Dev. Letts.*, **EDL-6**, p. 551, 1985.
33. P. K. Chatterjee, *Tech. Dig. IEDM*, 1979, p. 14.
34. M. L. Woods, "MOS VLSI Reliability and Yield Trends," *Proceedings IEEE*, Dec. 1986, p. 1715.
35. P. Yang and S. Aur, "Modelling of Device Lifetime due to Hot Carrier Effects," *Proc. of 1985 Internat. Symp. on VLSI Tech.*, p. 227.
36. C. Hu et al., *IEEE J. Solid-State Circuits*, **SC-20**, 1985, p. 295.
37. E. Takeda and N. Suzuki, "An Empirical Model for Device Degradation due to Hot Carrier Injection," *IEEE Trans. Electron Dev. Letts.*, **EDL-4**, p. 111, 1983.
38. S. Ogura et al., *IEEE Trans. Electron Dev.*, **ED-27**, 1980, p. 1359.
39. H. Mikoshiba, "Comparison of Drain Structures in n-Channel MOSFETs," *IEEE Trans. Electron Dev.*, Jan. 1986, p. 140.
40. L. C. Parrillo et al., *IEDM Tech. Dig.*, 1986, p. 244.
41. R.A. Chapman et al., *IEDM Tech. Digest*, 1987, p. 362.
42. Y. Tsunashima et al., *Proc. of VLSI Symposium*, 1985, p. 114.
43. T. Toyoshima et al., *Proc. of VLSI Symposium*, 1985, p. 362.
44. T. Noguchi, *Tech. Dig. IEDM*, 1986, p. 730.
45. S. Jain et al., *IEEE Electron Dev. Lett.*, Oct 1988, p. 539.
46. T. Y. Chan et al., *Electron Dev. Lett.*, Jan. 1986, p. 16.
47. J. R. Pfister and F.K. Baker, "Asymmetrical High Field Effects in Submicron MOSFETs", *EDM Tech. Dig.*, 1987, p. 51.
48. P. K. Ko et al., *Tech Dig. IEDM*, 1986, p. 292.
49. E. F. Da Silva, *Tech. Dig. IEDM*, 1987, p. 848.
50. M. -L. Chen, "Hot Carrier Aging in Two-Level Metal Processing," *Tech. Dig. IEDM*, 1987, p. 55.
51. F. C. Hsu, *Proc. of VLSI Symp.*, 1985, p. 108.
52. S. Fujita et al., *Tech. Dig. IEDM*, 1985, p. 64.
53. J. Y. Chen, "CMOS-The Emerging VLSI Technology," *IEEE Circuits and Devices Magazine*, March, 1986, p. 16.
54. M. Koyanagi et al., *IEEE Trans. Electron Dev.*, **ED-34**, 1987, p. 871.
55. C. Hu et al., *IEEE J. Solid-State Circuits*, **SC-20**, 1985, p. 295.
56. Y. Hiruta et al., *Tech. Dig. IEDM*, 1986, p. 718.
57. T. Mizuno et al., *Tech. Dig. IEDM*, p.726.
58. T. Y. Chan, J. Chen, P.K. Ko, and C. Hu, *IEDM Tech. Dig.* 1987, p. 718.
59. C. Chang and J. Lien, *IEDM Tech. Dig.* 1987, p. 714.
60. C. T. Sah, "The Evolution of the MOS Transistor," *Proceedings of the IEEE*, Oct. 1988, p. 1280.
61. P. Balk, "Effects of Hydrogen Annealing on Silicon Surfaces," presented at the *Electrochem. Soc. Spring Meeting*, May, 1965. Ext Abs. No. 109, p. 237.
62. M.-C. Jeng et al., "Design Guidelines for Deep-Submicrometer MOSFETs," *Tech. Dig. IEDM*, 1988, p. 386.
63. M. L. Chen et al., *Tech. Dig. IEDM*, 1988, p. 390.
64. R. W. Gregor, "Consequences of Ion Beam Shadowing in CMOS Source/Drain Formation," *IEEE Electron Dev. Letts.*, Dec. 1986, p. 677.

65. I. Ahmed, H. Naguib, and C. Gomez, *Ext. Abs. ECS Meeting*, Fall 1988, Abs. No. 282, p. 405.
66. S. H. Dhong and E. J. Petrillo, *J. Electrochem. Soc.*, Feb. 1986, p. 389.
67. F. Hsu and K. -Y. Chiu, *IEEE Electron Dev. Lett.*, EDL-5, p. 162, 1984.
68. K. - Y. Toh, P.-K. Ko, and R. G. Meyer, *IEEE J. Solid-State Ckts*, Aug. 1988, p. 950.
69. J. W. Brews et al., "Generalized Guide for MOSFET Miniaturization," *IEEE Electron Dev. Lett.*, EDL-1, p. 2, 1980.
70. R. H. Dennard et al., "Design of Ion-Implanted MOSFETs with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits*, SC-9, Oct. 1974, p. 256.
71. S. G. Chamberlain and S. Ramanan, *IEEE Trans. Electron Dev.*, Nov. 1986, p. 1745.
72. K. Marayam, J. C. Lee, and C. Hu, *IEEE Trans. Electron Dev.*, July 1987, p. 1509.
73. L. D. Yau, "A Simple Theory to Predict the Threshold Voltage in Short-Channel IGFETs," *Solid-State Electronics*, vol. 17, p. 1059, 1974.
74. P. Wright et al., *IEEE Electron Device Letts.*, August 1989, p. 347.
75. V. L. Rideout, F. H. Gaensslen, and A. LeBlanc, "Device Design Consideration for Ion-Implanted n -Channel MOSFETs," *IBM J. Res. Dev.*, p. 50, (1975).
76. R. Siquesch et al., *Tech. Dig. IEDM*, 1980, p. 429.
77. F. H. Hsu and K. Y. Chiu, *Tech. Dig. IEDM*, p. 96, 1984.
78. P. Wright et al., "Hot-Electron Immunity of SiO_2 Dielectrics with Fluorine Incorporation," *IEEE Electron Device Letters*, August 1989, p. 347.
79. Y. Nishioka et al., "Hot-Electron Hardened Si-Gate MOSFET Utilizing F Implantation," *IEEE Electron Device Letters*, April 1989, p. 141.
80. W. M. Werner, *Solid-State Electronics*, 17, 769, (1974).
81. R. R. Troutman, *IEEE J. Solid-State Circuits*, SC-9, 55 (April 1974).
82. J. R. Pfeister et al., *Tech. Dig. IEDM*, 1989, p. 781.
83. C.-Y. Lu et al., *IEEE Trans. Electron Dev.*, November, 1989, p. 2530.

PROBLEMS

- 5.1 (a) A p -channel MOS transistor is fabricated on an n substrate doped with 10^{15} phosphorus atoms per cm^3 , and a gate oxide which is 1000 Å thick. Calculate the threshold voltage if $\phi_{ms} = -0.6$ eV and $Q_{\text{tot}} = 5 \times 10^{11} \text{ cm}^{-2}$. (b) The threshold voltage of the MOSFET in part (a) is reduced by using ion implantation of boron. What is the required boron dose in order to obtain a threshold voltage of -1.5 V?
- 5.2 In the E-D NMOS process, depletion-mode NMOSFETs are used for load devices. This requires a negative threshold, which can be obtained by implanting a shallow arsenic or phosphorus dose into the channel region. Calculate the arsenic dose needed to achieve a -3 V threshold in a n^+ polysilicon-gate NMOS transistor which has a substrate doping of $3 \times 10^{16} / \text{cm}^3$ and a gate oxide thickness of 50 nm?
- 5.3 (a) Why is $\langle 100 \rangle$ -orientation preferred in NMOS fabrication? (b) What are the disadvantages if too thin a field oxide is used in NMOS devices? (c) How is a self-aligned gate obtained and what are its advantages? (d) List three functions of the PSG overglass layer?
- 5.4 In NMOS processing, the starting material is a p -type $10\text{-}\Omega\text{-cm}$, $\langle 100 \rangle$ -oriented silicon wafer. The source and drain are formed by arsenic implantation of 10^{16} ions/ cm^2 at

80 keV: the channel is implanted with 8×10^{11} boron ions/cm² at 30 keV through a gate oxide of 250 Å thickness. (a) Estimate the threshold voltage change caused by the boron ion implantation step. (b) Draw the doping profile along a coordinate perpendicular to the surface and passing through (b1) the channel region, and (b2) the source region.

5.5 Describe punchthrough current and subthreshold current and explain the differences between them.

5.6 Design a submicron MOSFET with a gate length of 0.75 μm. (The gate length is the channel length plus twice the junction depth.) If the junction depth is 0.2 μm, the gate oxide is 200 Å, and the maximum drain voltage is limited to 2.5 V, find the required channel doping so that the MOSFET can maintain its long-channel characteristics.

5.7 An *n*-channel MOSFET has a gate oxide 200 Å thick, an $L_{\text{eff}} = 1.0$ μm, source/drain junction depths of 0.2 μm, and a threshold voltage of 0.6 V. If the device is biased at $V_{\text{GS}} = 3$ V and $V_{\text{DS}} = 5$ V, calculate the saturation voltage V_{Dsat} , and the maximum electric field, E_{max} .

5.8 Explain the difference between the gate-threshold voltage, V_{T} , and the field-threshold voltage, V_{TF} , of a MOS device.

5.9 (a) Sketch a set of masks for fabricating an MOS transistor using Si-gate technology. (b) Repeat this exercise for an E-D NMOS inverter circuit with LOCOS isolation, buried contacts, and a depletion-mode load transistor.

5.10 An NMOS structure consists of an *n*-type substrate with $N_{\text{D}} = 5 \times 10^{15}$ cm⁻³, a gate oxide of 100 nm thickness, and an Al contact. The measured threshold voltage is -2.5 V. Calculate the surface charge density.

5.11 If a circuit designer wants to keep the off-state leakage current to less than 1 pA and also wants the threshold voltage value to be 0.9 V (where V_{T} is defined as the gate voltage that provides $I_{\text{DS}} = 1$ μA), what value of subthreshold swing (S.S.) will this device exhibit.

CHAPTER 6

CMOS PROCESS INTEGRATION

Complementary MOS (CMOS) is so-named because it uses both *p*- and *n*-type (complementary) MOS transistors in its circuits. (Figure 6-1 depicts a CMOS inverter.) Since CMOS technology is significantly more complex than NMOS with respect to the device physics and fabrication issues, the discussion here will include a thorough introduction to these subjects. (Note that an excellent comprehensive text on CMOS, dealing with both circuit and design issues, has recently been published.)¹

6.1 INTRODUCTION TO CMOS TECHNOLOGY

6.1.1 The Power-Dissipation Crisis of VLSI, and How CMOS Came to the Rescue

NMOS remained the dominant MOS technology as long as the integration level of devices on a chip was sufficiently low. It was inexpensive to fabricate, very functionally dense, and faster than PMOS. The earliest NMOS technologies required only 5 masking steps (including the pad mask).^{*} On the other hand, NMOS logic-gates (e.g., inverters) draw dc power during one of the inverter states. Therefore, an NMOS integrated circuit will draw a steady current even when being operated in the standby mode (i.e., even when no signal is being propagated through the circuit). Consequently, as the number of logic gates on the chip grows, the current being drawn (and hence the power being dissipated) also increases. Although this was always a limitation of NMOS for such applications as space-borne or portable electronic systems, it did not represent a drawback for most other applications when the number of devices on a chip was rela-

^{*} Early NMOS logic gates used butted contacts as well as enhancement-mode transistors for both the driver and the load. Depletion-mode loads and buried contacts came later. Hence, of the seven masks of the E-D NMOS process described in chapter 5, only five were required in early NMOS technology. This resulted not only in lower manufacturing costs per wafer, but also in higher yields. Thus, the cost of manufacturing NMOS circuits was brought even lower.

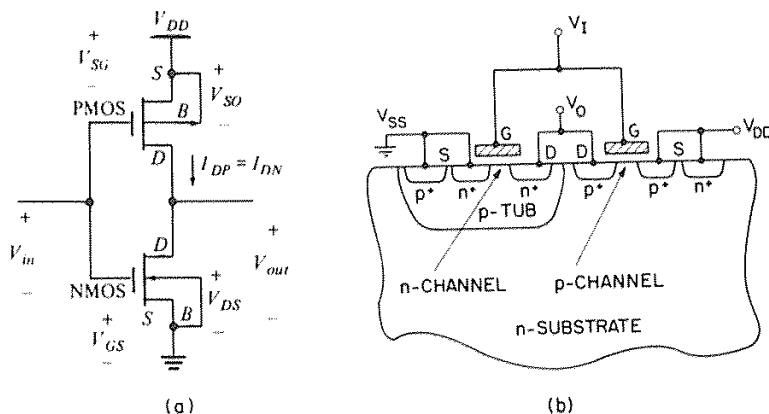


Fig. 6-1 CMOS inverter. (a) Circuit schematic. (b) Device cross section.

tively small. Such was the situation at the level of device integration that existed up to the mid-1970s.

With the dawning of the VLSI era, however, power consumption in NMOS circuits began to exceed tolerable limits. A lower-power technology was needed to exploit the VLSI fabrication techniques. CMOS represented just such a technology.

From a quantitative perspective, the ascendancy of CMOS was the inevitable result of the two-hundred-fold increase in functional density, and the twenty-fold increase in speed of integrated circuits between 1968 and 1987. For example, in 1969, 256-bit SRAM circuits (e.g., the Intel 1101) used 12- μm design rules to create the six-transistor SRAM cells; each cell occupied 20,600 μm^2 . In 1987, however, SRAM cells using 1.2- μm design rules occupied only 150 μm^2 . (The memory access-time in these respective memory circuits decreased from 1 μs to 46 ns.) By 1987 the decreased size and attendant increase in chip size allowed 256-kbit SRAMs to be built.

To take another example: The Intel 4004 4-bit microprocessor, introduced in 1971, had 2300 devices and was built in PMOS. The 8086 model, a 16-bit microprocessor introduced in 1978, had 29,000 devices and was built in NMOS, and the 80386, introduced in 1985, had 275,000 devices and was built in CMOS.

Chips can dissipate a maximum of about 5 W of power and still be used in conventional, but expensive, IC ceramic packages. In order for the much less expensive plastic packages to be used, however, the maximum power dissipation is limited to about 1 W. The Intel 8086 dissipated around 1.5 W of power when operated at 8 MHz. Thus, by the late 1970s it was already possible to manufacture NMOS chips whose power dissipation approached unacceptably high values. (Note that when the 8086 was later reissued in CMOS technology under the model number 80C86, its power consumption dropped to about 250 mW.)

In a CMOS inverter (unlike in an NMOS inverter) only one of the two transistors is driven at any one time. This means that when a CMOS inverter is not switching from one state to the other, a high-impedance path exists from the supply voltage to ground, regardless of the state the inverter is in. Hence, virtually no current flows, and almost no dc power is dissipated. CMOS thus allows the manufacture of circuits that need only several microwatts of standby power (Fig. 6-2).

The problem of power dissipation can also be considered from both a *chip* perspective and a *system* perspective. From the chip perspective, if microprocessors of the 32-bit generation were built in NMOS, they would dissipate 5 to 6 W of power. This would lead to severe heating and reliability concerns. In addition, expensive ceramic packages would be needed to house such chips. When such microprocessors are designed in CMOS, the power dissipation decreases to about 1 W.

From the system perspective, let's consider memory chips. Although a 1-Mbit DRAM may consume only 120 mW of power in NMOS, it consumes even less (~50 mW) in CMOS. Since there may be thousands of memory chips in a system (versus only a few microprocessors), the ramifications of lower power-dissipation are significant. Smaller power-supplies and smaller cooling fans are but two of these ramifications.

6.1.2 Historical Evolution of CMOS

Although CMOS is now the dominant integrated-circuit technology, for much of its 25-year history it was considered to be only a runner-up for the design of MOS ICs. The pairing of complementary *n*- and *p*-channel transistors to form low-power ICs was originally proposed by Sah and Wanlass in 1963.² The first CMOS ICs were fabricated

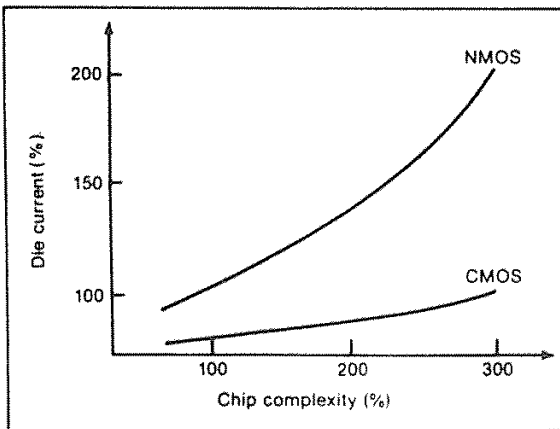


Fig. 6-2 As circuit complexity increases, NMOS power consumption rises to levels that eventually prevent further growth. In contrast, CMOS power consumption increases only slightly as the device count on a chip rises.

in 1966, and subsequent development of the technology was spearheaded by the RCA Corporation. The earliest volume commercial application was the use of the CMOS logic-inverter in the frequency divider circuits of digital watches.

CMOS technology at that time had many disadvantages compared to PMOS and, later, to NMOS. The drawbacks included significantly higher fabrication cost, slower speed, susceptibility to latchup, and much lower packing density. As a result, until the 1980s CMOS was limited to applications that could only be implemented with the technology's lower power dissipation (e.g., watches and calculators), or very-high noise margin (e.g., radiation-hard circuits). Furthermore, the advances made in NMOS fabrication were not rapidly transferred to CMOS, and for many years CMOS lagged behind the advanced Si-gate-NMOS and bipolar technologies. Except for the special applications mentioned above, it lay dormant for nearly a decade.

At the time CMOS circuits were first fabricated, the processes of ion implantation and local oxide isolation (LOCOS) had not yet been developed. In addition, metal gates were being used for MOS devices and control had not yet been established over the large and quite variable positive oxide charges in the gate oxide. As a result, *p*-well technology was the only means of fabricating CMOS.

While PMOS enhancement-mode transistors could be successfully fabricated in a lightly doped (e.g., 10^{15} cm^{-3}) *n*-substrate with a V_T of about -2 V, NMOS enhancement-mode transistors could not be fabricated on lightly doped *p*-substrates because the V_T of NMOS metal-gate transistors on such substrates is negative. In addition, the problems of oxide charge and the segregation of boron at the field-oxide/silicon-substrate interface, when combined with necessity of having to use relatively-thin field oxides (in order to be able to achieve adequate step coverage over nonrecessed field-oxide steps; see chap. 2), made it likely that parasitic channels would be established in the field regions between NMOS devices built on lightly doped substrates. Therefore, the only reliable way to manufacture enhancement-mode NMOS transistors for CMOS inverters was on regions with boron surface concentrations high enough to overcome these problems.

The *p*-well approach to building CMOS provided such regions, since the well had to be doped about ten times as heavily as the substrate for adequate control of doping in the well to be achieved. As a result, *p*-well technology became established in the companies that pioneered CMOS technology. Long after the problems of fabricating NMOS on lightly doped substrates had been solved (i.e., through control of the gate-oxide charge and of V_T by ion implantation) most of these companies continued to use *p*-well technology to design new circuits. While this was primarily due to historical inertia, *p*-well technology did have a few advantages over *n*-well technology (as will be discussed in section 6.2.2)

The packing-density limits of the early *p*-well CMOS technology, however, were responsible for its poor performance. The packing density was much worse than that of NMOS, primarily because the *n*- and *p*-devices had to be surrounded by guard rings (n^+ or p^+ diffusions that surround the device, as shown in Fig. 6-3)¹⁴ to prevent the inversion of the field regions.

Before LOCOS isolation and ion implantation became available, guard rings were needed to provide adequately large values of V_T in the field regions. However, their use

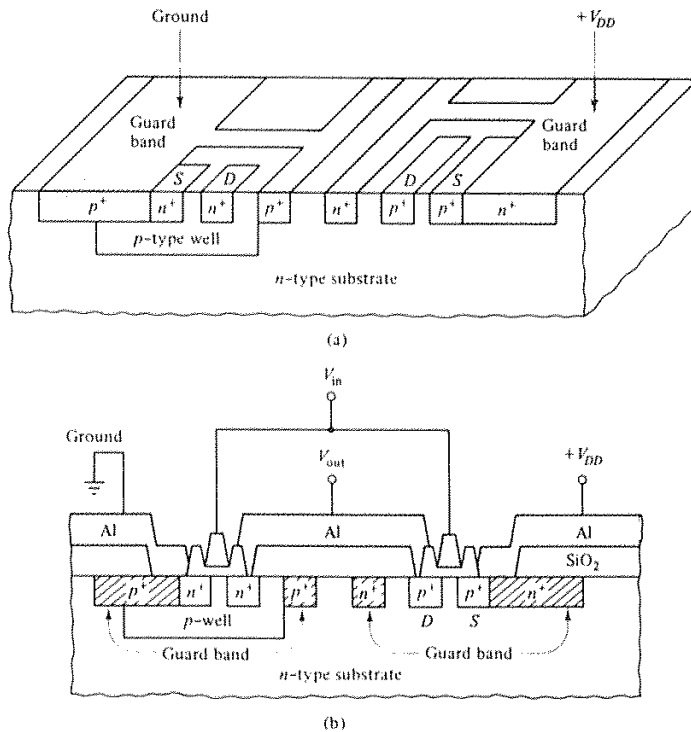


Fig. 6-3 CMOS structure with guard bands. A CMOS inverter circuit is depicted.¹¹⁴ From W. C. Till and J. T. Luxton, *Integrated Circuits: Materials, Devices, and Fabrication*. Copyright 1982 Prentice-Hall. Reprinted with permission.

results in very large-area devices.* As a result, until oxide isolation and channel-stop implants were developed, interconnect capacitance and resistance severely degraded the speed performance.[†] Once guard rings were no longer needed, however, the devices could be brought closer together, and the speed of CMOS circuits improved dramatically.

When it became apparent in the late 1970s that the increases in power density and dissipation would make it impossible to design future generations of MOS circuits in

* The drain regions must also be isolated from the guard rings by a lightly doped region to prevent avalanche or Zener breakdown, as seen in Fig. 6-3. This requirement exacerbates the area penalty when guard rings are used.

[†] However, it should be noted that the fabrication of NMOS devices in heavily doped *p*-wells also increased the device junction-capacitances and reduced the magnitude of the drive current. This further decreased the performance of the circuits, but to a lesser degree than did the interconnect capacitance.

NMOS, the companies that had been stubbornly continuing to use NMOS for design and fabrication finally began to consider CMOS. It was natural for them to seek a technology that was compatible with the modern high-production-volume Si-gate-NMOS processes that they had successfully perfected. Since n -well CMOS offers near compatibility with such processes, and since it allows n -channel transistor performance to be optimized (through fabrication in the lightly doped p -substrate regions), it became the technology of choice for many companies that had formerly been manufacturing NMOS integrated circuits.⁸

It became evident, however, that neither p -well nor n -well would be the optimum choice for submicron CMOS. Instead, it appeared that *twin-well CMOS* would be more effective. As a result, many such processes were subsequently developed.

We will examine the various well technologies with regard to their advantages and disadvantages in modern CMOS processes. Circuit designers and process engineers should be aware of the trade-offs. The advances developed through the refinement of NMOS were incorporated into CMOS technology, and the performance was thus dramatically improved over that exhibited by primitive CMOS circuits.

6.1.3 Operation of CMOS Inverters

The CMOS inverter (for which the circuit schematic and a sample layout are shown in Fig. 6-1a and 6-4, respectively) uses enhancement-mode transistors for both the NMOS driver and the PMOS load transistors. The gates of the two transistors are connected and serve as the input to the inverter. The common drains of each device are connected to the output of the inverter, and we assume that the inverter is driving some load capacitance, C_L (e.g., the input to another CMOS logic gate). Note that both the source and the body of the NMOS transistor are connected to ground, while those of the PMOS transistor are connected to V_{DD} (e.g., 5 V).

The threshold voltage of the NMOS driver transistor, V_{Tn} , is positive (e.g., $V_{Tn} = 0.8$ V), while that of the PMOS load transistor, V_{Tp} , is negative (e.g., $V_{Tp} = -0.8$ V). Figure 6-5a shows the inverter's output voltage, V_o , as a function of the input voltage V_{in} . This curve is known as the *static voltage input/output characteristic*, or the *transfer characteristic* of the gate.

When $V_{in} = 0$, $V_{GSn} = 0$ and the NMOS transistor is *OFF* (since V_{GSn} is < 0.8 V). The PMOS transistor, however, is *ON*, since $V_{GSp} = -5$ V, which is much more negative than -0.8 V. Thus, when $V_{in} = 0$, C_L is charged to V_{DD} through the turned-on PMOS load transistor, and $V_o = 5$ V.

When $V_{in} = 5$ V, the NMOS transistor is turned *ON*, since $V_{GSn} = 5$ V (which is > 0.8 V). The PMOS transistor is turned *OFF*, since $V_{GSp} = 0$ V (which is more positive than -0.8 V). Consequently, when $V_{in} = 5$ V (V_{DD}), the output is connected to ground through the turned-on NMOS driver, allowing C_L to be discharged. Since the PMOS device is off, C_L will be completely discharged, and V_o will be 0 V.

The most important property of the CMOS inverter is that when the gate is sitting quiescently in either logic state ($V_o = V_{DD}$ or 0), one of the transistors is *OFF* and the

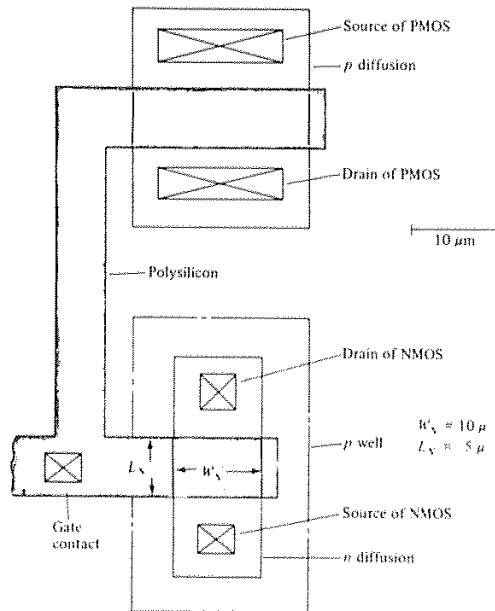


Fig. 6-4 CMOS inverter layout. From D. A. Hodges and H. G. Jackson, *Analysis and Design of Digital Integrated Circuits*, Copyright 1983, McGraw-Hill Book Co. Reprinted with permission.

current conducted between V_{DD} and ground is negligible (i.e., it is equal to the leakage current of the *OFF* device). This feature can be seen in Fig. 6-5b, which plots the current through the inverter, I_{DD} , as a function of V_{in} (solid curve). The power dissipated in the static (or standby) mode is then determined by the product of the leakage current and the supply voltage. Since the leakage current of an MOS transistor in cutoff (subthreshold current) is so small, very little power is consumed in the static mode. Another important feature is that the V_o swings all the way from V_{DD} to 0 as the inverter changes state; this characteristic of the output-voltage swing is referred to as swinging from *rail to rail*.

Two other operational features of the CMOS inverter must also be mentioned. First, as can be seen in Fig. 6-5b, significant current is conducted through the inverter only when both transistors are *ON* at the same time (i.e., during switching). Therefore, most of the power dissipation is due to the charging and discharging of C_L . In fact, it can be shown that the power dissipation is essentially given by $f C_L V_{DD}^2$, where f is the switching frequency.

Second, because the CMOS inverter output voltage can swing from rail to rail, it can inherently provide excellent noise margins. Noise margins are usually defined in terms of the *logic-gate threshold voltages*, V_{OH} , V_{OL} , V_{IH} , and V_{IL} (which are *not* the same parameters as *device threshold-voltages*, see Fig. 6-5a). The noise margins NM_L and NM_H (Fig. 6-5b) are defined according to the following equations:

$$NM_L = V_{IL} - V_{OL} \quad (6-1a)$$

$$NM_H = V_{OH} - V_{OL} \quad (6-1b)$$

Briefly, the argument for Eq. 6-1a is that the logic gate (e.g., *Inverter 1* in Fig. 6-6a) should provide a maximum "low output" that is less than the maximum "low input" that the subsequent logic gate (e.g., *Inverter 2*) can accept without causing the output voltage of *Inverter 2* to be driven into the ambiguous portion of the transfer characteristic. The difference between V_{IL} and V_{OL} (the noise margin) then specifies how much noise voltage can be tolerated at the input node of Inverter 2 before its output will be driven to a voltage value that is logically undefined.

In CMOS, V_{OL} approaches zero, and V_{OH} approaches V_{DD} . Because of the steep transition region in the transfer characteristic, V_{IL} and V_{IH} can be designed so that noise margins are on the order of $V_{DD}/4$ for expected process variations and operating temperatures. Thus, for a 5-V CMOS technology, V_{NML} can be 1.25 V, whereas in NMOS, V_{NML} is typically only 0.3 V.

For the performance of the logic gates to be maximized, the threshold voltages of the *p*- and *n*-channel transistors should be comparable (and, ideally, of equal magnitude). In addition, the threshold voltages should be as small as possible. The minimum values selected for V_{Tn} and V_{Tp} will be determined by the subthreshold leakage current, I_{Dst} .

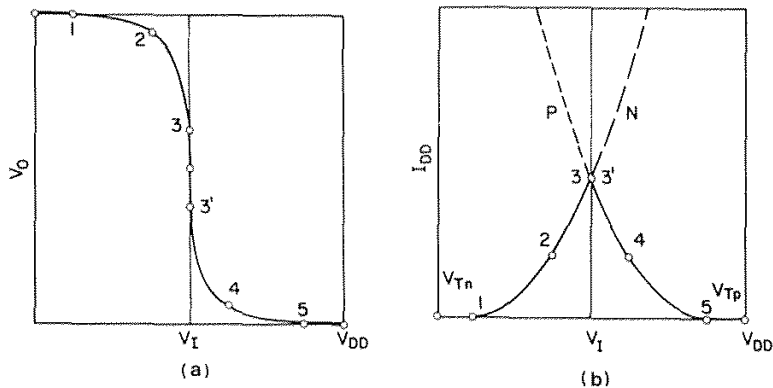


Fig. 6-5 (a) Output (V_O) versus input (V_I) voltage of CMOS inverter. (b) Current through inverter as a function of input voltage (solid curve); I-V characteristics of *n*- and *p*-channel transistors (dashed curves). The numbers correspond to different points on the inverter transfer characteristic.¹¹⁵

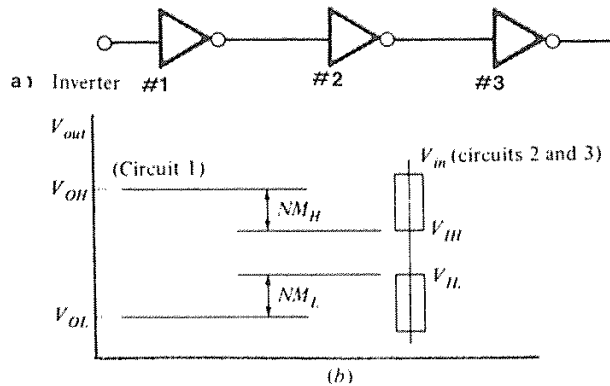


Fig. 6-6 (a) String of inverters connected in series. (b) Definition of noise margins.

For example, if the guideline described in the section on the subthreshold current in MOS devices is used (section 5.2.3 in chap. 5), a minimum value of V_{Tn} would be 0.5 V, to keep V_{OL} at least 0.5 V below V_{Tn} and to maintain sufficiently small I_{Dst} .

6.1.4 Advantages of CMOS

As noted earlier, the most important advantage of CMOS is its significantly reduced power density and dissipation.^{3,4,5,6} There are other advantages as well, which fall into the following main categories:

- device/chip performance
- reliability
- circuit design
- cost issues

These advantages, as well as the disadvantages of CMOS, will be discussed in this section.

6.1.4.1 Device/Chip Performance Advantages.

• Although logic gates designed in CMOS are larger than those of NMOS (primarily because of the increased spacing needed to isolate n -channel from p -channel devices), this packing density penalty is becoming less important, for several reasons.

First, the CMOS gate uses a PMOS device rather than a depletion-mode NMOS device as the load. The PMOS device is usually made about twice as wide as the NMOS device in order for symmetrical driving capability to be achieved. The NMOS depletion-mode load, however, is four times as large as the NMOS driver. Hence, the area of the two CMOS devices is actually smaller than the area of the two NMOS devices in an inverter.

Second, since the devices are built with submicron dimensions, the difference in the output drive capabilities (I_D) of identically sized PMOS and NMOS devices is decreased due to velocity saturation effects (Fig. 6-7a). The area of the CMOS inverter devices will thus grow proportionately even smaller.

Third, as devices become smaller, the fraction of the chip area required for interconnections becomes larger. Hence, the fact that the gate density is lower in CMOS in NMOS becomes less important. This is especially significant in random logic designs.

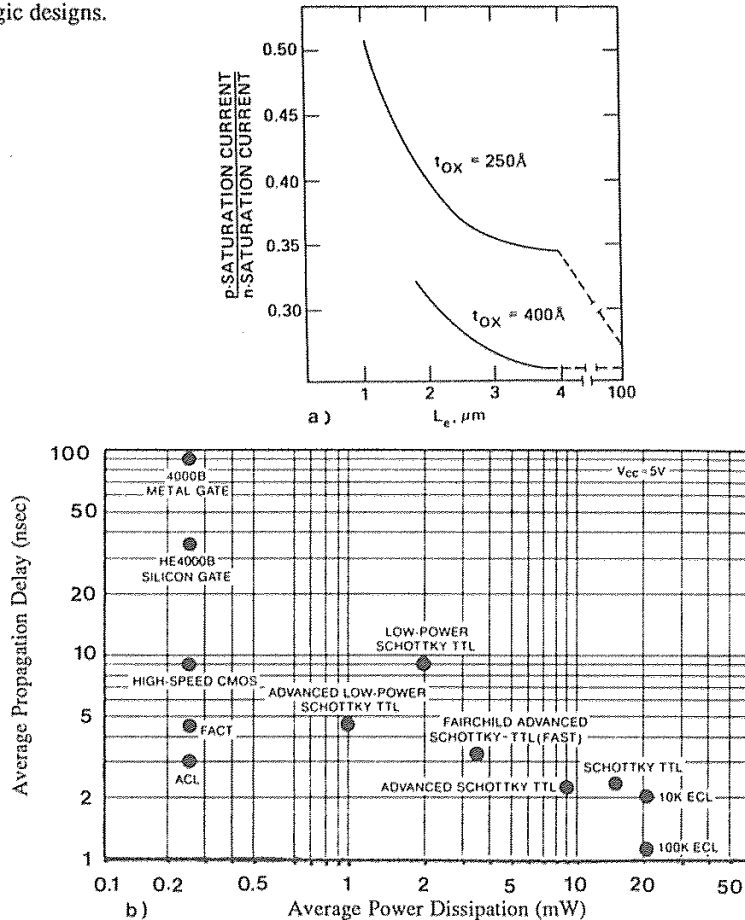


Fig. 6-7 (a) Ratio of p -channel saturation current to n -channel saturation current increases as the effective channel length of the devices shrinks, due to velocity saturation effects. (b) Comparative speed and power characteristics of various CMOS and bipolar logic families.

- Although the input capacitance of CMOS is larger than that of NMOS (since there are two MOS gates connected in parallel at the input of CMOS logic elements versus only one in NMOS), the interconnect capacitance is becoming more significant than the gate capacitance as gate sizes shrink and interconnect lengths grow with increasing chip size. NMOS gates are thus no longer significantly faster than CMOS gates. Furthermore, as the chip heats up from excessive power dissipation, the circuits slow down (due to degradation of carrier mobility). The lower power advantage might actually make a CMOS circuit operate faster than a comparable, but hotter-running, NMOS part.
- Since more devices can be placed onto a single chip in CMOS than in NMOS, less chip-to-chip driving is necessary, and the overall system speed is improved.
- The fan-out capacitance for MOS devices is much smaller than that of bipolar transistors. Therefore, while the transconductance is also much smaller, if the load capacitance (other than fan-out) is small, much less current is needed to charge the next MOS gate compared to a bipolar logic gate. As the channel length of MOS device shrinks, it may be possible to decrease the delay time required to charge other gates on the same chip. On the other hand, because the performance of bipolar devices is relatively insensitive to shrinking device dimensions, the delay times in CMOS are approaching those of bipolar ECL, but at much lower levels of power dissipation (Fig. 6-7b).
- CMOS can be operated over a wider range of V_{DD} values (e.g., 2-7 V) than NMOS.
- Threshold body-biasing sensitivity is less important in CMOS, and bootstrapping is not needed for transference of a signal through a string of inverter.
- The radiation hardness of CMOS circuits is much higher than that of NMOS.

6.1.4.2 Reliability Advantages.

- The heat generated during operation can raise the chip temperature to the point at which it becomes more prone to failure. Since CMOS circuits dissipate much less power, in most cases they should be inherently more reliable.
- Hot-carrier degradation of MOS devices should be decreased in CMOS for several reasons. First, since hot-electron effects are much less severe in PMOS devices, the load devices in CMOS will suffer less degradation than the depletion-mode load devices used in NMOS. Second, the CMOS gates do not draw static current, so long-term, cumulative hot-electron induced degradation will be smaller. Finally, unlike NMOS, CMOS generally does not use bootstrapping (which increases the electric field in the device and thus aggravates hot-electron degradation).
- Electromigration failures in the metal lines of the circuit are reduced, since no static current flows in the metal lines.

- The soft error rates (SERs) of DRAMs and SRAMs can be reduced by one to two orders of magnitude when the memory arrays are fabricated inside a well region where doping type is the opposite of the substrate's. The added SER protection arises because the reverse-biased well-substrate junction of CMOS creates a potential barrier against carriers generated in the substrate.

6.1.4.3 Circuit-Design Advantages.

- CMOS can achieve "static ratioless" logic design. Circuits that contain a p -channel transistor for every n -channel transistor are said to be "static," because such gates are triggered by the data path signal and do not require the use of an external clock. The design is said to be "ratioless" because the inverter voltage transfer characteristic does not depend strongly on the relative geometric sizes of the p and n transistors. By contrast, commonly used "ratioed" NMOS circuitry must have transistor widths and lengths that are chosen both to balance the currents between transistors and to ensure that this balance is maintained, given changes in operating temperatures, power-supply voltage variation, and day-to-day differences in the manufacturing process.
- As noted earlier, since the output of CMOS logic gates swings from rail to rail, excellent noise margins are inherently provided.
- The flexibility in selecting transistor geometry provided by the ratioless nature of CMOS makes it much easier for "uncommitted circuit" designs, such as gate arrays and standard cells, to be implemented. In gate arrays and standard cells, the number of input and output signals (fan-ins and fan-outs) are normally not known when the individual transistors are laid out. As a result, it is very difficult with PMOS or NMOS to create gate-array and standard-cell configurations that possess sufficient design flexibility. Instead, most standard cells and virtually all MOS gate arrays are implemented in CMOS.
- In NMOS, gating (or pass) transistors reduce the transmitted signal by the so-called *threshold loss*, whereas in CMOS such transmission gates leave the signal unchanged. As a result, signal regeneration can be less frequent in clocked CMOS circuits.
- CMOS allows both analog and digital functions with high circuit densities to be implemented on the same chip. This has stimulated the development of switched-capacitor techniques for analog-to-digital conversion and allowed their integration on a chip with a digital-signal algorithm.
- The circuit-design benefits of CMOS for analog applications are that the switches have no offset voltage, and that the area required for operational amplifiers is much smaller than that needed for NMOS op amps. That is, while an NMOS op amp might take 30 transistors, the same op amp in CMOS might take one-third the number of transistors, as well as one-third the area. Furthermore, CMOS op amps are three to five times smaller than bipolar op amps.

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 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 $2\ \mu\text{m}$ (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 $1\ \mu\text{m}$ 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

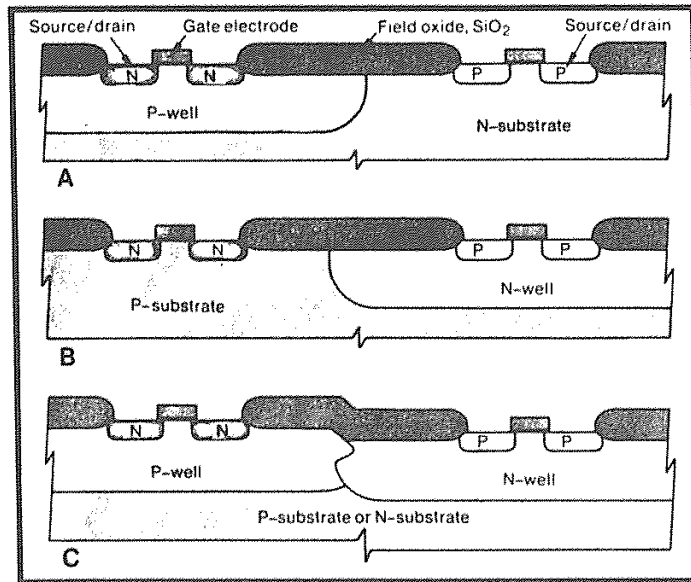


Fig. 6-8 CMOS circuits can be designed using (a) *p*-well, (b) *n*-well, or (c) twin-well technology.³ (© 1983 IEEE).

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 3×10^{14} - $1 \times 10^{15}/\text{cm}^3$ 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 *n**p**n* bipolar transistor as an on-chip driver or *n*-channel FETs for an analog input). Third, it is less

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 $V_{DD} = 5\text{ V}$. The substrate doping of the p -type wafers is $1 \times 10^{15}/\text{cm}^3$. The n -wells are to have an average dopant concentration of $1 \times 10^{16}/\text{cm}^3$. The p -channel MOSFET sources and drains are to have junction depths of 0.4 μm and an average dopant density of $10^{18}/\text{cm}^3$. 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 $V_{DD} = 5\text{ V}$ (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).

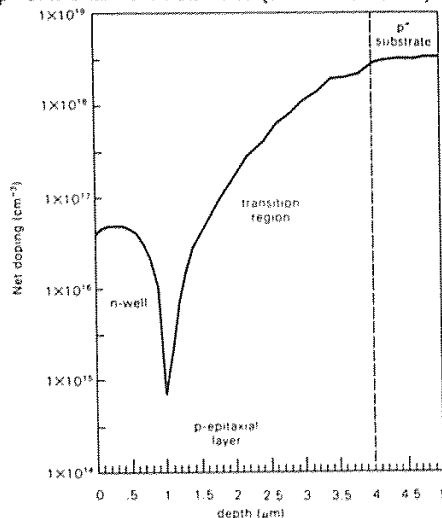


Fig. 6-9 A p -on- p -epi doping profile and diffused n -well profile measured after the entire CMOS process has been completed. The as grown epi thickness is 4 μm and the final lightly doped epi thickness is about 2 μm .¹¹⁶ (© 1987 IEEE).

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_0 , is very small (i.e., $k_{0,\text{Sb}} = 0.023$; 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

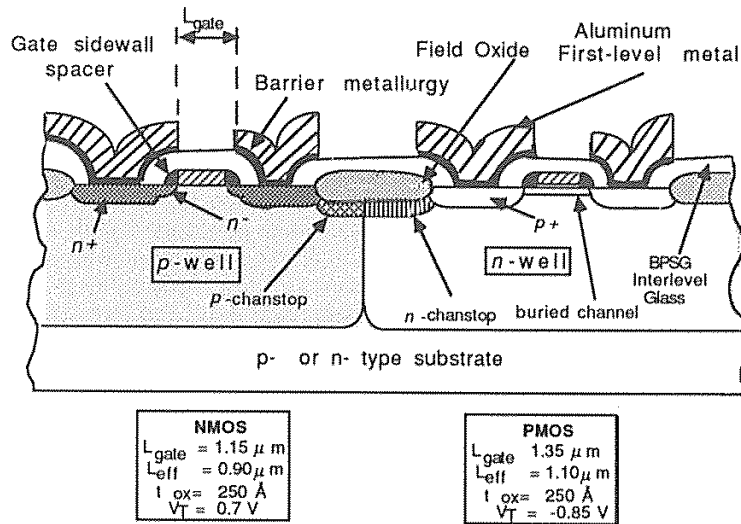


Fig. 6-10 Cross section of a twin-well 1.2- μm CMOS structure.¹² Reprinted with permission of Semiconductor International.

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 $\Omega\text{-cm}$ (7×10^{14} boron/ cm^3) epi on a 0.005 $\Omega\text{-cm}$ (2×10^{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 $\Omega\text{-cm}$ 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

a heavily doped substrate. In either case, the level of surface doping is significantly lower than that required for building either the p - or n -channel MOSFETs. Each of the well dopants is implanted separately into the lightly doped surface region and is then driven in to the desired depth.

The doping profiles of each of the device types can be set independently, since the constraint of single-well CMOS does not exist (i.e., that the well doping must always be higher than the doping of the substrate in which one type of device is made). This was originally cited as an advantage of twin-well CMOS over single-well CMOS, with the argument made that both device types could thus be optimized.¹⁴ This claim for 1-2 μm CMOS has been questioned by Chen,⁴ who points out that in modern single-well CMOS processes, an additional implant is used to prevent punchthrough without the need to raise the entire substrate-doping concentration. By incorporating this additional implant, it is possible to build higher-performance devices than can be achieved with the twin-well approach, in terms of junction capacitance and sensitivity to body effect.

Twin-well CMOS does offer some significant benefits for devices with submicron-channel lengths (although these advantages are gained at the cost of greater process complexity).

The first, and most important advantage arises when devices with submicron channel lengths are fabricated. Since it is recognized that the two device types perform similarly as channel lengths approach 0.5 μm , it is useful to provide symmetrical n - and p -channel devices. Furthermore, at submicron dimensions the body doping of both transistor types must be raised significantly to prevent punchthrough and to maintain adequate threshold-voltage levels. Thus, the advantage of having one type of MOS transistor in a lightly doped region (to optimize its performance at the expense of the other) disappears. It is instead more beneficial to produce two types of active device wells, each with its own optimized doping profile (i.e., formed by separate implants into a lightly doped substrate).*

The second advantage of the twin-well process is that it is compatible with the technologies of either isolation by selective epitaxial growth (SEG) or trench isolation.⁴ Both approaches restrict the lateral diffusion of the dopants in each of the wells. In addition, sidewall inversion along the trench is less likely when both device sidewalls are butted against a trench that has been formed in a more highly doped well.¹⁵ Finally, when deep trenches are used with a thin epitaxial layer on a heavily doped substrate, latchup can be eliminated. The combined use of the twin-well process and advanced isolation techniques allows n^+ to p^+ spacing to be dramatically reduced in comparison to single-well technologies.

* It is not useful to start with a substrate doped to the optimum level needed for just *one* of the submicron devices. If this were done, a single well of much higher doping would have to be established for the other type of submicron device, which would unnecessarily degrade the device performance. It is possible, with the twin-well approach, to have both types placed in a well of optimum doping profile.

A third benefit is that the combination of epitaxial substrates and the twin-well process allows either substrate type to be chosen with no effects on transistor performance and essentially no change in process flow. Such flexibility is useful since some applications are best met with n^+ substrates and others with p^+ substrates. This advantage may be important if a single process is needed for implementing designs with different applications.

Finally, self-aligned channel stops can be easily implemented in the twin-well approach, allowing the spacing between n - and p -channel devices to be reduced. Although this spacing reduction is not as great as it would be if trenches or SEG were used, the process is much less complex with the twin-well method.

6.2.6 Retrograde-Well CMOS

Conventional wells are formed in single- and twin-well CMOS technology by implanting dopants and then diffusing them to the desired depth. However, the diffusion occurs laterally as well as vertically, which has the effect of reducing packing density. If a high-energy implant is used to place the dopants at the desired depth without further diffusion, much less lateral spread occurs (see Vol. 1, chap. 9). Such high-energy implants also cause the peak of the implant to be buried at a certain depth within the silicon substrate (depending on the implantation energy), and the impurity concentration decreases as it approaches the wafer surface. Since the well profile in this case is different from that of conventionally formed wells (in which the doping concentration is highest at the surface, and decreases monotonically with depth), such deeply implanted wells are known as *retrograde wells*. The retrograde-doping profile is retained by minimizing the temperature cycles of later process steps. Retrograde wells can be implemented on both bulk wafers and on epitaxial wafers.⁵

Besides the potential benefit of increased packing density, such wells offer the following advantages:⁴

- A retarded electric field is created in the parasitic vertical bipolar transistor (thereby providing some protection against latchup; see section 6.4).
- Susceptibility to vertical punchthrough is reduced.
- The conductivity in the bottom of the well is increased, which also provides some further latchup protection (as will be explained in section 6.4.5).
- A higher threshold voltage can be achieved in the field regions of the p -wells since the boron implant is done following field oxidation (i.e., the boron does not segregate out into the field oxide as it is grown).¹⁶
- Lateral diffusion of the boron is also eliminated, thereby reducing encroachment of the boron into the active regions.

A disadvantage of the retrograde-well approach is that both the junction capacitance and body factor are significantly increased. For example, in a simulation of a 32-bit CMOS arithmetic logic unit (ALU), it was found that the circuit delay increases by

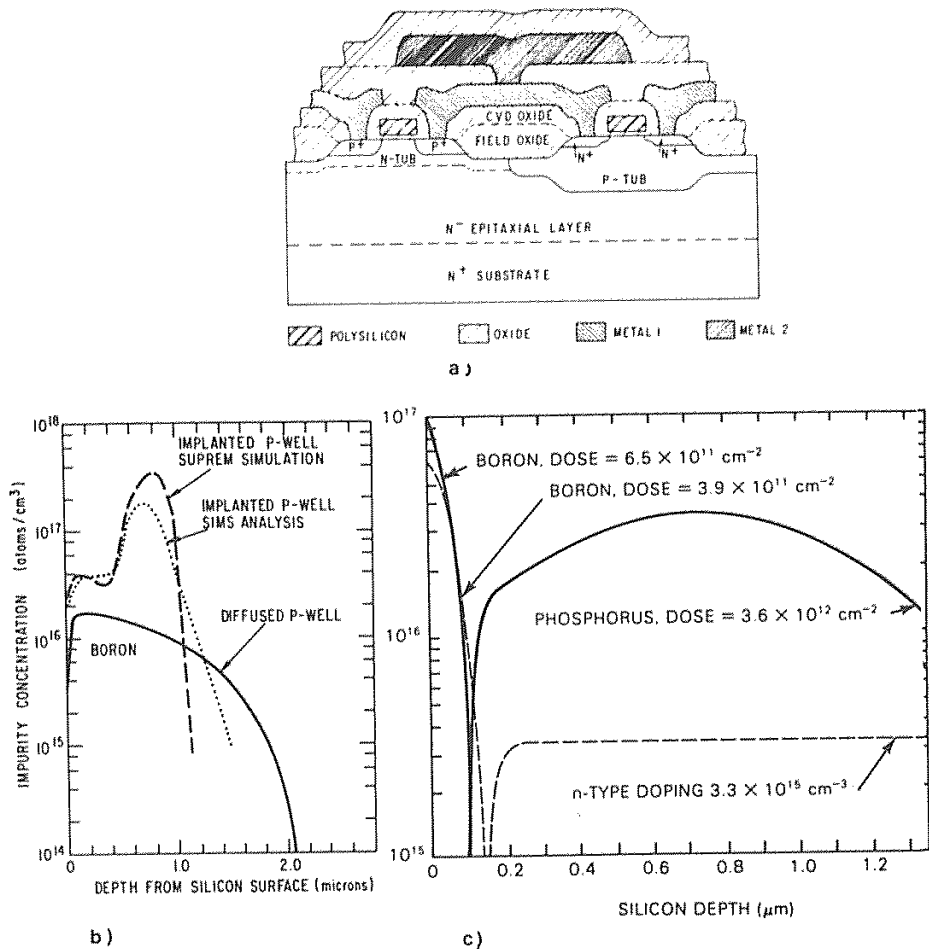


Fig. 6-11 (a) Twin-tub device cross section in which a retrograde *p*-well is used. General Electric AVLSI Process. (b) Retrograded *p*-well implanted impurity concentration profile. General Electric AVLSI process. Also shown is a conventionally thermally diffused well.⁵ (© 1986 IEEE). (c) Simulated doping profiles in PMOS channel regions: dashed lines, lightly doped substrate; solid lines, retrograde *n*-well.²³ (© 1982 IEEE).

about 7% as the *p*-channel junction capacitance increases by 30 percent.⁴ When the retrograde well is formed by means of a very high-energy implant, however, the doping concentration under the bottom of the source and drain regions is reduced, which reduces

the junction capacitance. This implies that if a very-high-energy ion implanter (~ 1 MeV) were available as a manufacturing tool, the disadvantage of retrograde wells could be overcome.¹⁷ In addition, the higher doping that will be required in the wells for fabrication of submicron CMOS devices may mean a less severe junction capacitance penalty.

Although both p -type¹⁸ and n -type¹⁹ retrograde wells have been demonstrated (see Figs. 6-11a and b,⁵ and Fig. 6-11c,²³ respectively), the p -type technology has been more widely implemented. The reason for this is that a 700-keV (or greater) ion implanter is required for the formation of n -type retrograde wells. As of 1989 such machines were commercially available (e.g. NV 1003), but not yet found in ordinary production environments because of their relatively high cost. The p -type retrograde well, on the other hand, can be formed either by implanting singly ionized boron at 400 keV or doubly ionized boron at 200 keV (because of the larger projected range of boron compared with arsenic or phosphorus; see Vol. 1, chap. 9). Although the doubly ionized boron approach is achievable with conventional production implanters, it is not a trivial process to implement. In addition, a quadruple-well technology has been developed. This approach uses deep retrograde p - and n -wells, as well as shallow p - and n -wells (Fig. 6-12).¹¹²

A twin-retrograde-well 0.7- μm -CMOS process for fabricating 1-Mbit SRAMs has been reported.⁷⁴ The high energy implants also allow a restricted thermal budget to be used, thus reducing the up-diffusion from the p^+ substrate. This permits the use of a thinner epi layer (which also helps prevent latch-up), and also allows implantation of

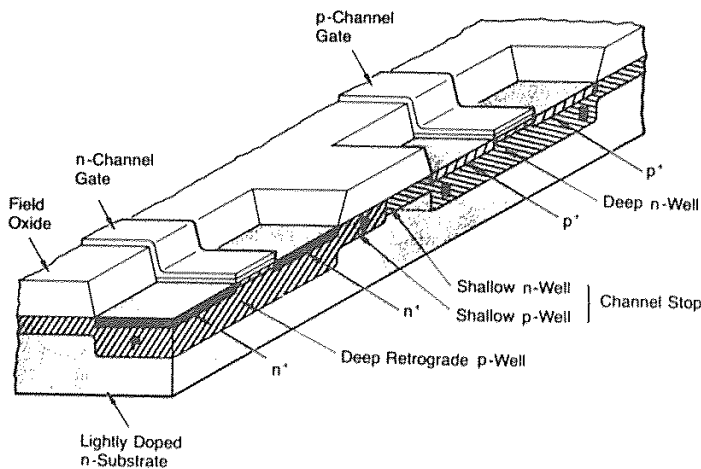


Fig. 6-12 Cross section of a quad-well CMOS device.¹¹⁷ Reprinted with permission of VLSI Design.

the channel-stop dopants after the field oxide has been grown and planarized (thus minimizing lateral diffusion of the boron channel-stop dopants).

6.2.7 Summary of CMOS Well-Technology Issues

The selection of a particular CMOS process depends largely upon circuit applications, and to a lesser degree on technology evolution. For VLSI circuits with 1-2 μm design rules, n -well CMOS has been the more widely used single-well technology, both because n -channel devices can be optimized in the p -substrate and because most circuits use more NMOS than PMOS devices. In circuits that use transistor channel lengths smaller than 1 μm , twin-well and retrograde-well technology will become more attractive. As noted, the price will be increased process complexity.

The optimum process-integration design strategy is to first select the well technology (and to decide whether or not to use epitaxial substrates), and to then select the isolation method. Once these decisions have been made, the well depth and doping profile can be determined. The well depth will impact the lateral-diffusion distance of the well and the vertical-punchthrough voltage. The doping profile will affect transconductance, threshold voltage, source/drain punchthrough, junction capacitance, carrier mobility, source/drain-to-substrate breakdown, sensitivity to body effect, and hot-electron effects.

6.3 p -CHANNEL DEVICES IN CMOS

The fabrication of p -channel devices in CMOS presents some unique problems which arise from the need to build both NMOS and PMOS devices on the same chip. The problems revolve around the choice of a doping type for the polysilicon gate electrode and the impact of this choice on the threshold voltage and transistor action of PMOS devices.

6.3.1 PMOS Devices with n^+ -Polysilicon Gates

As mentioned earlier, the threshold voltages of the n - and p -channel devices in a CMOS circuit should have comparable magnitudes for optimal logic-gate performance. To allow for maximum current-driving capability, the threshold voltages should also be as small as possible, with the minimum value dictated by the need to prevent excessive subthreshold currents under normal circuit operating conditions. For 5-V CMOS technologies, typical threshold voltages are 0.8 V for V_{Tn} and -0.8 V for V_{Tp} . Furthermore, the most common choice for the gate material is heavily doped n -type polysilicon. The work function of n^+ polysilicon is ideal for n -channel devices since these will yield threshold voltages of less than 0.7 V for reasonable values of channel doping and oxide thicknesses.*

* Note that the polysilicon layer may be combined with a layer of silicide for sheet resistance reduction; since the polysilicon remains as the underlayer of the polycide, the work function of the gate electrode will not be changed.

Figure 5-4, chapter 5 shows the value of V_T in devices manufactured with n^+ polysilicon gates (left scale)¹⁸ as a function of doping and various gate-oxide thicknesses (Q_{tot} is assumed to be small enough that its effect on V_T can be ignored). The value of V_{Tn} is less than 1.0 V for gate-oxide thicknesses of 25 nm or less and a substrate doping of less than 10^{17} cm^{-3} . Thus, the threshold voltage of NMOS devices can easily be adjusted to the desired value of 0.7 V by means of ion implantation.

When n^+ polysilicon is used for the gate electrode of PMOS devices, however, it is not as easy to adjust V_{Tp} to -0.7 V. Figure 5-4 shows V_{Tp} (on the left scale for n^+ poly) as a function of substrate doping and gate-oxide thickness. In the doping range of 10^{15} - 10^{17} cm^{-3} , V_{Tp} is already more negative than -0.7 V. Thus, implanting the n -doped body with more n -type dopant would only raise the *magnitude* of V_{Tp} , rather than bringing its value closer to the desired -0.7 V. To reduce the magnitude of V_T in PMOS devices using an n^+ polysilicon gate, it is necessary to implant the channel with a shallow layer of boron. The dose must be heavy enough to overcompensate the n -surface so that a p -region depleted of holes is formed. This shifts V_{Tp} toward more positive values by forming a compensating layer.

The fact that boron is implanted to adjust both V_{Tn} and V_{Tp} in CMOS circuits with n^+ polysilicon gates suggests that a single implant could be used instead of two separate implants. Figure 6-13 shows that this can be accomplished if the appropriate background dopings are chosen for the substrate and the well.²¹ On the other hand, it

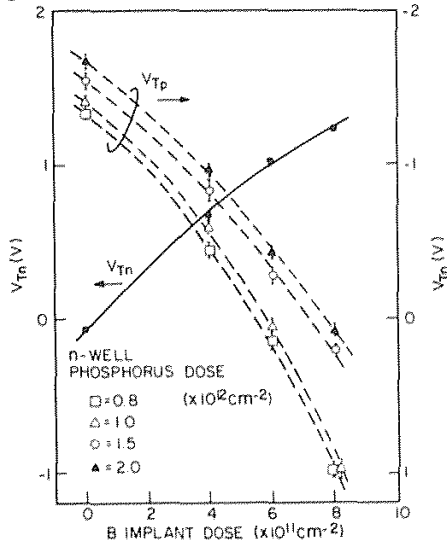


Fig. 6-13 Threshold voltages of n -channel (V_{Tn}) and p -channel (V_{Tp}) transistors as a function of boron threshold-adjustment dose. The CMOS structure uses an n -well implanted into a p -substrate (whose doping level is $6 \times 10^{14} \text{ atoms/cm}^3$). V_{Tp} results are shown for various implant doses of the n -well.²¹ (© 1980 IEEE).

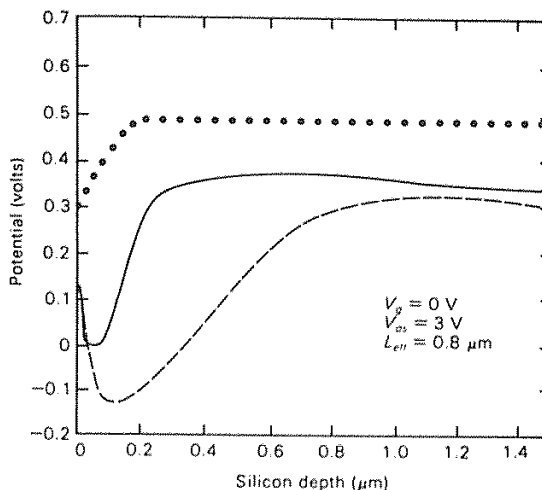


Fig. 6-14 Channel potential profiles in PMOS device in CMOS. *Dashed lines:* buried channel PMOS in a lightly doped substrate. *Solid lines:* buried-channel PMOS in a retrograde *n*-well. *Dotted lines:* surface channel NMOS drawn for comparison.²³ (© 1982 IEEE).

may be decided to use two separate implants in order to achieve better short channel behavior through individual optimization of the *n*- and *p*-channel devices.

6.3.1.1 Punchthrough Susceptibility. PMOS devices in which boron is used to adjust V_T exhibit a high susceptibility to punchthrough effects, since the boron implant produces a *p*-layer with a finite thickness. The potential minimum in the channel is thus moved away from the Si-SiO₂ interface (Fig. 6-14), causing more current to flow below the surface of the device.²³ Such PMOS devices are referred to as *buried-channel transistors*. As seen in Fig. 6-14 (which gives the *calculated* variation of the potential as a function of distance below the surface), the potential minimum moves further into the substrate as the thickness of the implanted *p*-layer is increased.

As the potential minimum moves deeper below the surface, the punchthrough susceptibility also becomes more pronounced (see section 5.5.2). This is illustrated in Fig. 6-15, which shows the results of a simulation²² in which the lines of equipotential in the channel region were plotted for various depths of the channel junction Y_j (for a constant source/drain junction depth of 0.15 μm). As Y_j is increased from 0.1 to 0.2 μm , the drain electric field extends closer to the source for a constant gate and drain bias. Hence, this simulation predicts that more barrier lowering will occur as Y_j is increased, leading to an increase in the punchthrough current.

The calculated predictions are supported by experimental data, as shown in Fig. 6-16, which plots the subthreshold I_D - V_{GS} characteristics of the structures described in Fig. 6-15. The subthreshold swing (S.S.) has the smallest slope when Y_j is 0.2 μm (indicating that the largest punchthrough current flows in this case). In fact, when

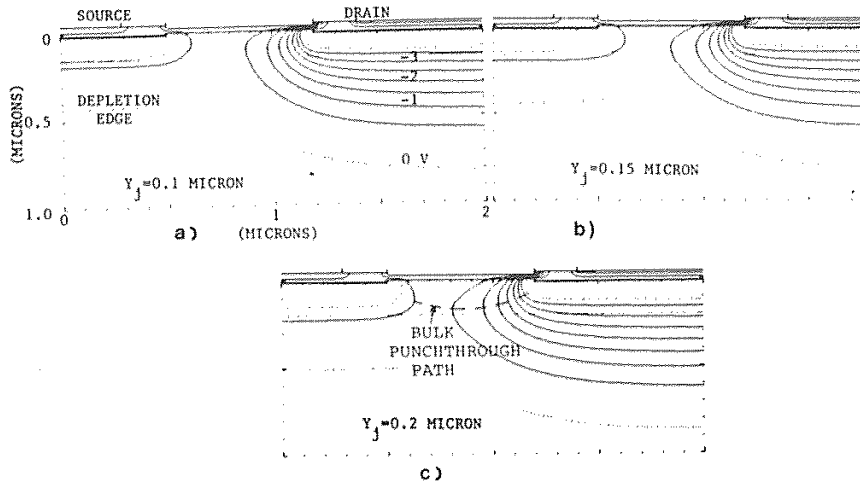


Fig. 6-15 Simulated 2-D potential profile for p -channel MOSFETs with $Y_j = 0.1, 0.15$, and $0.2 \mu\text{m}$. The drain and gate bias are -3 V and 0 V , respectively.²² From K. M. Cham, S.-Y. Oh, D. Chin, and J. L. Moll, *Computer-Aided Design and VLSI Device Development*, 2nd Ed., Copyright Kluwer Academic, 1989. Reprinted with permission.

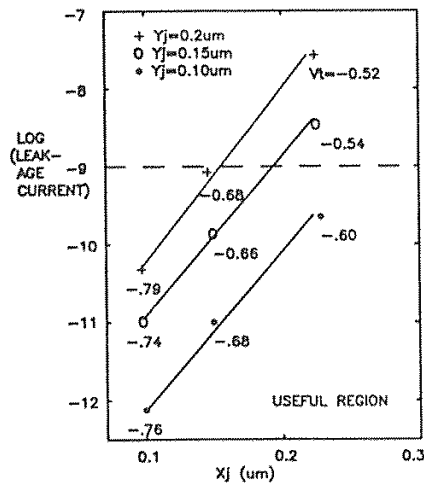


Fig. 6-16 Simulated subthreshold leakage current versus Y_j and X_j . $L_{\text{eff}} = 0.5 \mu\text{m}$, $W = 50 \mu\text{m}$, $t_{\text{ox}} = 25 \text{ nm}$, $V_{\text{GS}} = 0 \text{ V}$, and $V_{\text{DS}} = -3 \text{ V}$.²² From K. M. Cham, S.-Y. Oh, D. Chin, and J. L. Moll, *Computer-Aided Design and VLSI Device Development*, 2nd Ed., Copyright Kluwer Academic, 1989. Reprinted with permission.

$V_{GS} = 0$, I_{Dst} is increased by two orders of magnitude as Y_j is increased from 0.1 to 0.2 μm .

Leakage currents due to punchthrough in PMOS devices can be a significant problem in some CMOS IC applications. For ULSI devices, even a small value of leakage current per device may not be tolerable. For example, it has been shown that in submicron PMOS transistors ($L_{eff} = 0.8 \mu\text{m}$), the punchthrough current will increase by two orders of magnitude if V_{DS} is increased from -1 V to -10 V.²⁴ Thus, if the punchthrough current were 1 nA at -1 V, it would be increased to 0.1 μA if V_{DS} were increased to -10 V. Such leakage would cause the dissipation of a few-tenths of a watt of power in a chip containing 1 million PMOS transistors.

The most obvious solution is to increase the PMOS device channel length. This is done in many CMOS technologies, and it is the reason that the minimum channel lengths of PMOS devices are often larger than those of the NMOS devices on the same chip. Another obvious technique is to make the p -buried layer as thin as possible. One of the reported approaches for achieving this involves implantation with BF_2^+ (which produces shallower boron layers than implantation with boron, see Vol. 1, chap. 9). Another approach is to use a high-energy n -implant (e.g., As at 400 keV) in order to place more n -type dopant atoms below the pn junctions (the more heavily doped regions absorb the drain voltage in a shorter distance, while simultaneously squeezing the channel pn junction toward the surface).²⁵

To prevent a shallow implanted-boron layer from growing thicker, it is necessary to use a reduced thermal budget in order to restrict the process sequence following the implant in order to restrict boron diffusion. Specific steps for restricting boron redistribution include the following:¹²

1. Implant the boron through the gate oxide. This avoids the oxidation-enhanced diffusion of boron that would occur during the growth of the gate oxide if the implant were performed first. In this case, it is necessary to prevent the gate oxide from becoming contaminated during implant, either by material sputtered by the beam line or by vaporized resist material used as a mask against the implant. To prevent such contamination, a thin layer of polysilicon may be deposited on the gate oxide prior to the implant (in fact, immediately after the oxide is grown).²⁶
2. After the implant has been performed, the remainder of the polysilicon film is deposited. This polysilicon is doped during the deposition step (at $\sim 600^\circ\text{C}$) in situ with phosphorus, in order to avoid the 900°C phosphorus doping thermal cycle that would have to be used if the poly were doped following deposition.
3. A BPSG glass layer is used as the dielectric between the gate and the first level of metal. A significantly lower temperature can be used to flow BPSG than PSG (e.g., 850°C versus 1000°C). A lower temperature cycle can thus be used to smooth the surface topography (flow step) and gently taper the contact holes after etch (reflow step).

The use of a boron implant to adjust V_{Tp} becomes less feasible as devices use even thinner gate oxides, since larger doses of boron are needed. Y_j thus becomes deeper, and the punchthrough problem worsens. Solutions involving the use of gate electrodes other than n^+ polysilicon must therefore be explored. One alternative is to use p^+ polysilicon for PMOS devices; another is to use a material whose work function is very close to the mid-gap of silicon, thereby allowing symmetrical threshold voltages for n - and p -channel devices.⁵

6.3.2 PMOS Devices with p^+ -Polysilicon Gates

When p^+ polysilicon is used for the gate material, V_{Tp} is shifted from the values that occur when n^+ polysilicon is used. Figure 5-4, chap. 5 plots V_{Tp} as a function of substrate doping when p^+ polysilicon is employed (using the scale on the right side of the figure). V_{Tp} can be seen to be less negative than -0.7 V over the substrate doping range of 10^{15} - 10^{17} cm^{-3} . Thus, it can easily be made more negative through the implantation of phosphorus or arsenic into the channel. If p^+ polysilicon is used throughout the circuit, however, the NMOS device then has to be overcompensated in order for V_{Tn} to be reduced to sufficiently small values. (Note that some reports have been published on the use of p^+ polysilicon alone.^{27,28}) This implies that it would be ideal to use both n^+ and p^+ poly gates on the same chip (with n^+ poly for NMOS devices, and p^+ poly for PMOS devices).¹⁰⁰

Such a dual-doped poly approach, however, would add process complexity, and would also introduce other problems arising from the need to connect the two types of poly (e.g., at the input of an inverter). Such problems occur when a silicide overlayer (or *strap*) is used to make the connection between n^+ and p^+ poly (as a method to avoid a separate, space-consuming metal contact). Because the silicide strap provides a very rapid diffusion path for boron and arsenic, one type of poly can be counterdoped by the other when the device is subjected to high-temperature excursions. This counterdoping can occur to the degree that a region of poly can change doping types (i.e., from n^+ to p^+). In such cases, the threshold voltage of devices with counterdoped poly will be shifted from their designed value.

If the processing temperature is limited to 800°C after the two types of poly have been connected by the silicide, such lateral diffusion does not produce significant shifts in V_T .²⁹ On the other hand, temperatures of 900°C will produce sufficient counterdoping to significantly shift V_T . Since one of the last high temperature steps in CMOS is activation of the source/drain implants, this would mean either using a lower-temperature activation step, or performing the step prior to formation of the silicide layer. Process considerations for 0.5- μm CMOS technologies using both n^+ and p^+ -poly gates are described in reference 85.

Another approach is to form a polysilicon electrode with an overlying conductor layer that suppresses such counterdoping. One such structure is a W-TiN-poly electrode structure.¹⁰⁵ The thin (30-nm) TiN film acts as a diffusion barrier to the dopants in the poly and also prevents reaction between the poly and the W. Very little of the dopant

present in the poly is found to diffuse into the W, even following an anneal at 900°C for 1 hour.

Another problem encountered with p^+ poly gates when a thin gate oxide is used is poor V_T process control, due to penetration of the boron into the oxide (or further, into the silicon).⁷⁸ It is reported that boron will penetrate gate oxides that are ≤ 12.5 nm thick during a 900°C 30-minute post-implant anneal in N_2 .⁸¹ This implies that if p^+ poly gates are used, a lower processing temperature may be needed.⁸⁵

If too low a temperature is used, however, the boron implanted into the polysilicon will not be sufficiently redistributed. The polysilicon dopant concentration at the polysilicon–gate oxide interface could thus end up being less than the mid- $10^{19}/\text{cm}^3$. This would make the work function of the polysilicon different from the desired degenerately doped value, creating V_T control problems in the MOS devices.

It has also been found that the presence of fluorine in the gate oxide worsens the boron penetration problem (Fig. 6-17a).^{123, 124} Such fluorine can be introduced into the gate oxide if the PMOS source drain regions are implanted using BF_2 . Elemental boron is therefore considered inherently superior to BF_2 as the implant species for surface-channel PMOS devices in CMOS technologies that use p -doped polysilicon. A study of enhanced boron diffusion through thin SiO_2 layers in a wet oxygen atmosphere is also reported in reference 107.

6.3.4 Gate Materials with Symmetrical Work Functions (for Both NMOS and PMOS Devices)

Because the larger work functions of molybdenum (4.7 V), tungsten, or refractory silicides produce low and nearly symmetrical threshold voltages for both PMOS and NMOS devices on moderately doped Si substrates, work has been conducted to investigate their suitability as gate-electrode materials.³⁰ For example, $TaSi_2$ gates have been successfully implemented.³¹ Some of the benefits that such gate materials provide (besides symmetrical threshold-voltage values) include a reduction in subthreshold leakage currents and a decreased sensitivity to body bias (Fig. 6-17b).

Molybdenum and tungsten films deposited by means of magnetron sputtering have also been evaluated in terms of adhesion to SiO_2 , mechanical stress, and compatibility with silicon processing techniques.^{32,33} It was found that both could be deposited with low compressive stress by adjusting the deposition conditions of the sputter process. Both films can also exhibit good adhesion to SiO_2 . (An additional advantage of Mo and W is that their resistivities are about 100 times lower than that of doped polysilicon and about 10 times lower than that of polycide gates.)

Some novel techniques had to be developed in order for compatibility to be established with the conventional Si-gate MOS process sequences. For example, it is necessary to use a wet H_2 ambient to oxidize the silicon without oxidizing the Mo^{33} or W gates.³⁰ (This procedure is useful when a screen oxide is to be formed prior to the source/drain implant, but after the formation of the gate sidewall spacers used in LDD structure.) In addition, Mo and W form a layer of columnar grains, which makes such films susceptible to ion implant channeling along the grain boundaries. When the

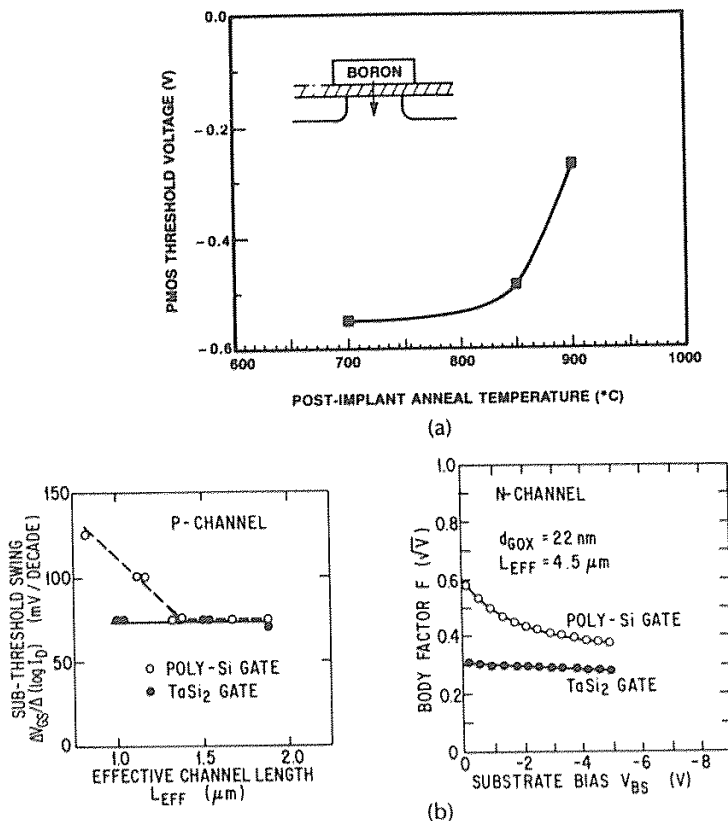


Fig. 6-17 (a) Boron penetration through the gate oxide causes the threshold voltage in BF₂-implanted PMOS devices to shift positive at anneal temperatures above ~800°C.¹²³ (© 1989 IEEE). (b) Sub-threshold characteristics versus effective channel length for poly-gate and TaSi₂-gate PMOS devices.¹¹⁸ (© 1984 IEEE).

source/drain implantation is performed, some of the dopants may thus penetrate the gate film. This problem can be overcome by limiting the ion energy³³ or by depositing a thin layer of PSG (~60 nm thick) on the gate electrodes prior to implantation.³⁰

Finally, V_T control with refractory metal gates was once a problem because of the relatively high concentrations of radioactive impurities (e.g., U and Th) in Mo and W sputtering targets. Such impurities can also produce soft errors in large memories. Recently developed chemical-purification procedures have significantly reduced the concentrations of U and Th in these sputtering target materials.³⁴ Methods for depositing Mo³⁵ or W⁹⁸ by CVD offer another option for forming of high-purity films. A deep-submicron CMOS process that uses a W gate has been reported.⁷⁷

6.4 LATCHUP IN CMOS

A major problem in CMOS circuits is the inherent, self-destructive phenomenon known as *latchup*.³⁶ Latchup is a phenomenon that establishes a very low-resistance path between the V_{DD} and V_{SS} power lines, allowing large currents to flow through the circuit. This can cause the circuit to cease functioning or even to destroy itself (due to heat damage caused by high power dissipation).

The susceptibility to latchup arises from the presence of complementary parasitic bipolar transistors structures, which result from the fabrication of the complementary MOS devices in CMOS structures. Since they are in close proximity to one another, the complementary bipolar structures can interact electrically to form device structures which behave like *pnpn* diodes. In the absence of triggering currents, such diodes act as reverse-biased junctions and do not conduct. It is possible, however, for triggering currents to be established in a variety of ways during abnormal (but nevertheless frequently occurring) circuit-operation conditions (e.g., terminal overvoltage stress, transient displacement currents, ionizing radiation, or impact ionization by hot electrons). Since there are many such parasitic *pnpn* structures on a VLSI CMOS chip, it is possible to trigger any one of them into latchup.

The phenomenon of latchup is well understood,³⁷ and many approaches have been implemented to control or even eliminate it. However, since the problem is increasing in severity as device dimensions continue to shrink, new latchup suppression techniques will be needed.

Latchup is fairly complex and can be a difficult concept to grasp for readers not well-versed in device physics. We will therefore briefly review the device concepts relevant to the problem before discussing the processing, layout, and circuit design techniques that have been developed to solve it.

6.4.1 Parasitic *pnpn* Structures in CMOS Circuits

Our example device configuration will be a *p*-well CMOS technology. As shown in Fig. 6-18a, lateral *pnp* and vertical *npn* transistor structures are inevitably created in *p*-well CMOS as a result of the multiple diffusions needed to fabricate *p*- and *n*-channel devices. The emitter of the lateral *pnp* transistor is the p^+ source and/or drain, while its base is the *n*-substrate and its collector the *p*-well. The n^+ source and/or drain comprises the emitter of the vertical *npn* device, while the *p*-well forms its base and the *n*-substrate its collector. The equivalent resistances of the substrate (R_{sub}) and the well (R_w) are also important elements of the latchup structure.

Several major aspects of these parasitic devices can be seen by studying the CMOS device cross-sections shown in Figs. 6-18a and 6-18b. First, the base of the *npn* transistor is connected to the collector of the *pnp* transistor (i.e., they are part of the same *n* region in the CMOS structure), and the base of the *pnp* is connected to the collector of the *npn*. We can therefore draw a simplified equivalent circuit diagram of this connection (Fig. 6-18c). It can be seen that the base of each transistor is driven by the collector current of the other, forming a positive feedback loop. Next, we note that

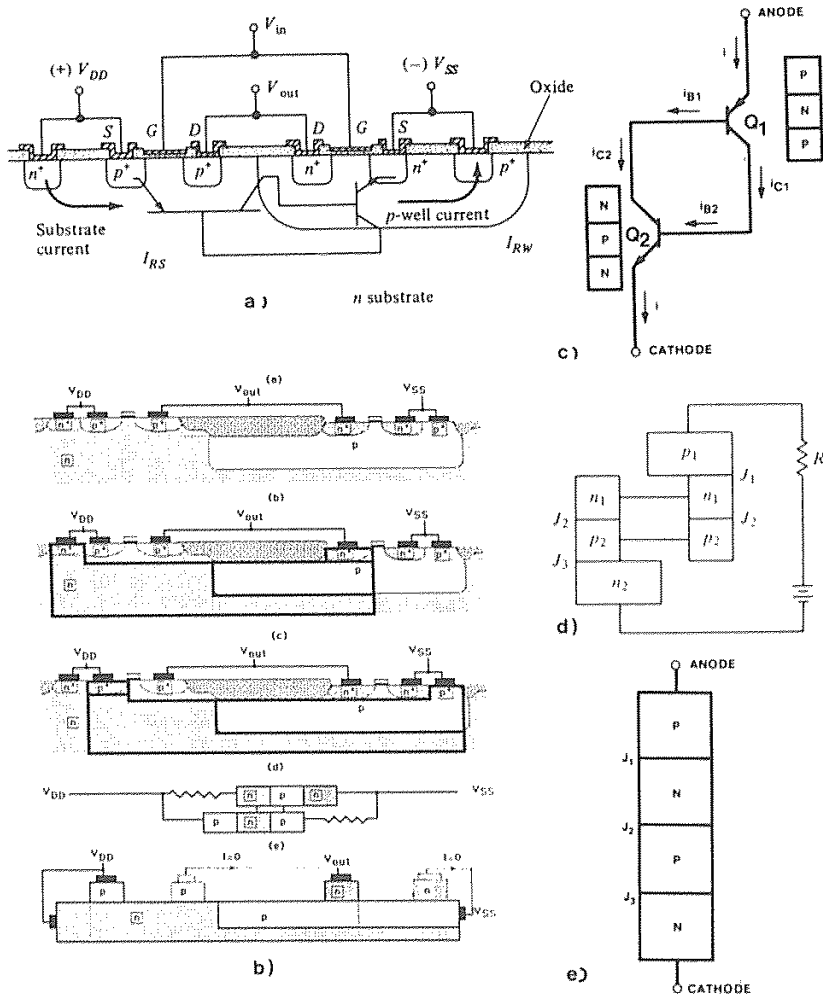


Fig. 6-18 (a) Cross-sectional view of a p -well CMOS inverter with parasitic bipolar transistors and lateral currents schematically shown.⁴ (© 1986 IEEE). (b) Schematic showing the parasitic nnp and pnp transistors in a p -well CMOS inverter. From W. Maly, *Atlas of IC Technologies*, Copyright 1987 by the Benjamin/Cummings Publishing Company. Reprinted with permission. (c) Simplified equivalent circuit diagram showing how these parasitic transistors are connected in the CMOS structure. (d) Schematic version of the circuit diagram of part (c). (e) pnp diode obtained by merging the connected n_1 and p_2 regions shown in part (d).

the emitter of the *nnp* is connected to V_{SS} (e.g., 0 V), and the emitter of the *pnp* is connected to V_{DD} . These external power-supply connections can also be added to the equivalent circuit diagram.

It should be noted that this simplified diagram ignores the parasitic resistances of the electrical paths through the bulk regions of the silicon, R_{sub} and R_w . These resistances, however, *are* very important in the understanding and control of latchup, and they will therefore be incorporated as soon as the simplified description of the device behavior is presented (Fig. 6-18c). The simple *pnpn* diode circuit to be used in the initial discussion turns out to be a "worst-case" possibility. That is, the addition of R_{sub} and R_w to the circuit model actually reduces the latchup susceptibility of CMOS. Since the two transistors are connected via their base and collector regions, by merging both the n_1 and the p_2 regions in Fig. 6-18d, we come up with a parasitic device structure (Fig. 6-18e).

6.4.2 Circuit Behavior of *pnpn* Diodes

Devices with structures like those shown in Fig. 6-18e and with external connections to the two end regions only, are known as *pnpn diodes*.* The terminal connected to the p_1 region is called the *anode*, and the terminal connected to the n_2 region is called the *cathode*. When an external voltage is applied with the anode voltage positive with respect to the cathode, and the resulting current I is measured, the I - V characteristic of this device is observed to have four distinct regions (Fig. 6-19); as follows:

Region 1. For voltages with values from a to b , very little current is observed to flow from anode to cathode, and the device is said to be in an *OFF*, *forward-blocking*, or *high-impedance* state. In this state, junctions J_1 and J_3 are forward-biased, and J_2 is reverse-biased. The externally applied voltage appears primarily across the reverse-bias junction.

Region 2. For voltages with values from b to c , the voltage across the reverse-bias junction, J_2 , approaches the breakdown voltage of that junction. The current during this voltage excursion increases slowly up to the breakover voltage, V_{BO} (point c), at which point it suddenly increases abruptly.

Region 3. For voltages from c to d , the device exhibits a differential negative resistance (i.e., the current increases as the voltage sharply decreases). This is a transient state, which occurs as the device switches from the *OFF* state to *Region 4* operation (*ON* state).

Region 4. For voltages beyond c , the I - V characteristic exhibits *low-imped-*

* Note that the generic term *diode* simply means a device with two electrodes. Therefore, we should not expect the electrical behavior of all two terminal electronic devices to be alike. More specifically, we should not confuse *pnpn diodes* with *pn diodes*, with respect to electrical behavior. In fact, as we shall see, the I - V characteristic of the three junction *pnpn* diode is considerably different, and more complex, than that of the single-junction *pn* diode.

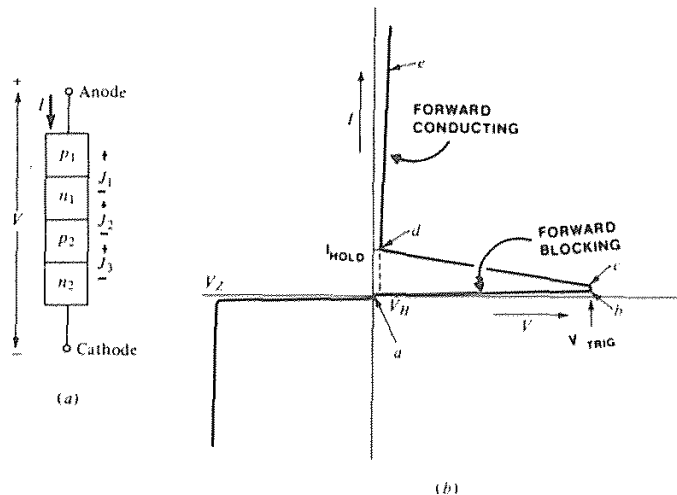


Fig. 6-19 (a) *pnpn* diode showing current reference direction and voltage polarities used in part b. (b) I-V characteristic of a *pnpn* diode.

ance behavior. The I-V curve in this region behaves very much like that of a single forward-biased *pn* junction and is known as the *ON*, (or *forward conducting*) state. In this state the junction J₂ is forward-biased, and the voltage across the entire device is on the order of 1.0 V. If the current through the device is reduced while it is operated in this state, the device will remain *ON* only as long as I exceeds I_H (the so-called *holding current*). If the current is reduced below I_H , the diode switches back to the high-impedance state. The voltage across the device when I_H is flowing is called the *holding voltage*, V_H . Thus, a *pnpn* diode operated with a positive voltage between the anode and cathode is a bistable device that can switch from an *OFF* state to an *ON* state, or vice versa. If the external circuit can supply more current than I_H , the device will remain *latched* in the *ON* state as long as power is applied.

The parasitic *pnpn* structure in the CMOS circuits exhibits essentially the same I-V characteristic as the device just described. If the parasitic *pnpn* diode is triggered into operating in Region 4, and if the external circuit can provide the necessary holding current, the CMOS circuit will remain *latched up* in the *ON* state, even if the source of trigger current has been removed.

6.4.3 Device Physics Behavior of *pnpn* Diodes

Assume that a voltage source through a resistor once again applies a positive voltage to the *pnpn* diode, as shown in Fig. 6-19a, and that this produces a current I through the device. In *Region 1* the applied voltage forward-biases J₁ and J₃ and reverse-biases J₂.

If the emitter-base junction of a bipolar transistor is forward biased and the collector-base junction is reverse-biased, the transistor is in the *active* mode of operation. Thus, in the equivalent circuit of Fig. 6-18c, transistors Q_1 and Q_2 are both biased in the active mode when the diode is operated in *Region 1*. If a bipolar transistor is operated in the active region, the collector current of a transistor is given by $I_C = \alpha I_E + I_{CO}$. When this equation is applied to Q_1 and Q_2 , then

$$I_{C1} = -\alpha_1 I + I_{CO1} \quad (6-2a)$$

and

$$I_{C2} = -\alpha_2 I + I_{CO2} \quad (6-2b)$$

where α_1 and α_2 are the common-base current gains of Q_1 and Q_2 , respectively.

According to Kirchhoffs current law, the sum of the currents entering Q_2 must be zero. Using this law and the current components as shown in Fig. 6-18c, we get

$$I - I_{C1} - I_{C2} = 0 \quad (6-3)$$

Combining Eqs. 6-2 and 6-3, we obtain

$$I = \frac{(I_{CO2} + I_{CO1})}{1 - (\alpha_1 + \alpha_2)} \quad (6-4)$$

It can be seen that as the sum of $\alpha_1 + \alpha_2$ approaches unity, the current I increases without limit. At this point the device is said to *break over*.

In bipolar transistors, the magnitude of α increases with collector current at low current levels (as shown in Fig. 6-20a). Since the value of I_C is increased as the avalanche breakdown voltage of the collector junction is approached, an increase in collector voltage can therefore lead to a significantly increase α .

When the collector currents are small, both α_1 and α_2 are much less than 1, and the current flowing through the diode is essentially the sum of the leakage currents, $I_{CO} = I_{CO2} + I_{CO1}$. As a result of the effects that lead to the increase of α , however, if the voltage across the *pnpn* diode is increased to a point near the collector-base junction breakdown voltage, the magnitude of the two alphas also increases. If the sum of the two alphas approaches unity, the current I begins to rise rapidly, which further increases the magnitudes of the alphas (this is an example of a *positive-feedback*, or *regenerative*, mechanism). When $\alpha_1 + \alpha_2 = 1$, breakover occurs, and *Region 3* operation sets in.

Beyond this point, device stability is provided by forward-biasing of the junction J_2 . Since both junctions in each of the two transistors are now forward-biased, both Q_1 and Q_2 are in the *saturation* region of operation. In saturation, the current gain of a bipolar transistor α again becomes smaller. As a result, once the diode enters the *ON* region, the transistors enter saturation to the degree necessary to maintain the condition of $\alpha_1 + \alpha_2 = 1$. The current I then increases to a value that is essentially limited by the external circuitry.

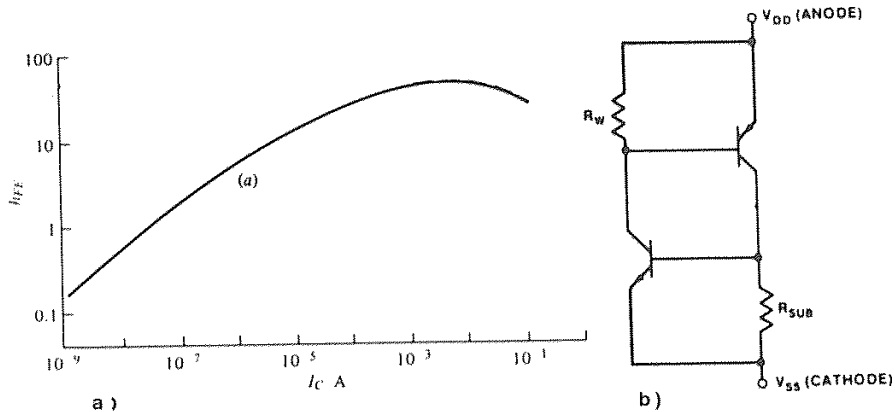


Fig. 6-20 (a) Dependence of current gain on the collector current. (b) Equivalent circuit model of the latchup structure in CMOS, including the parasitic resistances, R_W and R_{SUB} .

In the *ON* region, the voltage across the device is the algebraic sum of the voltages across the three forward-biased junctions. Since the voltage drop across the center junction J_2 is in the opposite direction of the voltage drops across J_1 and J_3 , the total drop across the *pnpn* diode in the *ON* region is about 1.0 V.

To maintain the diode in the *ON* region, the condition of $\alpha_1 + \alpha_2 = 1$ must continue to be satisfied. The holding current is the minimum current at which this condition is still met. If the current through the device is reduced to less than I_H (or if the applied voltage drops below V_H), the diode switches back to the *OFF* region of operation.

We can also express the condition $\alpha_1 + \alpha_2 = 1$ in terms of the β of the device by adding $(\alpha_1 \alpha_2)$ to both sides of the equation and rearranging terms, to get

$$\alpha_1 \alpha_2 = \frac{(\alpha_1 \alpha_2)}{1 - \alpha_1 - \alpha_2 + \alpha_1 \alpha_2} \quad (6-5)$$

This can be further rearranged to give us

$$\frac{\alpha_1}{(1 - \alpha_1)} \frac{\alpha_2}{(1 - \alpha_2)} = 1 \quad (6-6)$$

Since $(\alpha_1 / 1 - \alpha_1) = \beta_1$, and $(\alpha_2 / 1 - \alpha_2) = \beta_2$, we finally get

$$\alpha_1 + \alpha_2 = \beta_1 \beta_2 = 1 \quad (6-7)$$

We will refer to the expression $\beta_1 \beta_2 = 1$ as the *current-gain-product latchup condition for an ideal pnpn diode*.

6.4.4 Summary of Necessary Conditions for Latchup

Based on the principles of operation of *pnpn* diodes, the following four conditions must exist in a CMOS circuit in order for latchup to occur:

Latchup Condition #1. The emitter-base junctions of both parasitic bipolar transistors must be forward-biased. (It commonly happens, however, that initially only one transistor has its emitter-base junction forward-biased. This condition may then supply the current necessary to forward-bias the emitter-base junction of the second bipolar transistor.)

Latchup Condition #2. The product of the transistor gains must be sufficiently large to allow regenerative feedback. (We will see that the minimum product of β_{npn} and β_{pnp} needed to induce latchup in real CMOS structures is usually much greater than 1. That is, the $\beta_{npn}\beta_{pnp}$ product needed to cause latchup = 1 in ideal *pnpn* devices, but >1 in actual CMOS structures.)

Latchup Condition #3. The external circuit must be able to supply a voltage equal to or greater than the holding voltage, V_H , of the *pnpn* structure.

Latchup Condition #4. There is a minimum latchup trigger time for which the stimulus must be present in order for the circuit to be latched.

6.4.5 Circuit Behavior of *pnpn* Structures in CMOS Circuits

In actual CMOS circuits, the parasitic *pnpn* device structures consist of two complementary parasitic bipolar transistors adjacent each other. The parasitic series resistances of the electrical path from the *n*-well contact to the *nnp* collector, R_w , and of the path from the substrate contact to the *pnp* collector, R_{sub} , are also important circuit elements of the structure. Hence, the equivalent circuit that most accurately models the latchup structure in CMOS must include R_{sub} and R_w as well as the two parasitic bipolar transistors (Fig. 6-20b).

One of the effects caused by each of these two resistors is that a portion of the collector current of each transistor is siphoned away, reducing the base current. As a result of such current shunting, the effective current gains of the bipolar transistors are reduced. In fact, when finite R_{sub} and R_w are included in the circuit, Eq. 6-7 – in terms of the current gains β_{npn} and β_{pnp} – must be modified to more accurately express *Latchup Condition #2* for *pnpn* structures in CMOS circuits. For the latchup structures in actual CMOS, therefore, the inequality is expressed as

$$\beta_{npn}\beta_{pnp} > 1 + \frac{(\beta_{pnp} + 1) \left(I_{R_{sub}} + \frac{I_{R_w}}{\beta_{pnp}} \right)}{I - I_{R_{sub}} - I_{R_w} \left(1 + \frac{1}{\beta_{pnp}} \right)} \quad (6-8)$$

where $I_{R_{sub}}$ and I_{R_w} are the currents that flow in R_{sub} and R_w , respectively. The minimum gain-product needed to induce latchup in CMOS circuits can thus be much greater than 1, depending on the values of the series resistances.

In general, the approaches to eliminating latchup can be divided into two categories: (1) those that reduce the bipolar transistor current gains, and (2) those that lower the value of the series resistances R_{sub} and R_w . If either of these approaches is successful, the latchup condition given by Eq. 6-8 will be harder to satisfy. In the extreme, for example, if R_{sub} and R_w could be made equal to zero, the two emitters would be short-circuited and would thus be prevented from ever turning on. Even if R_{sub} and R_w are only made smaller, the values of $I_{R_{sub}}$ and I_{R_w} will increase, making the right side of the inequality larger. Therefore, various techniques have been developed to decrease the values of R_{sub} and R_w .

6.4.5.1 Value of β in CMOS Vertical Parasitic Bipolar Transistors.

The current gain of the vertical parasitic bipolar transistor in an actual CMOS structure depends on the well depth, the well doping concentration, and the built-in field in the well (see chap. 7). In typical 1- μm CMOS structures, the well is 1-2 μm deep and is doped to $\sim 1 \times 10^{16} \text{ cm}^{-3}$. In n -well technology, the current gain of the vertical pnp transistor, β_{pnp} , will be ~ 100 . In p -well or twin-well technology, the current gain of the vertical nnp transistor, β_{nnp} , will be two to three times larger than that of the pnp device in n -well technology (due to the higher minority-carrier mobility in the base region).

The transient response of the bipolar transistors is also important because of the minimum latchup trigger time (*Latchup Condition #4*). In addition, latchup in a real circuit is normally induced by transient triggering. The transit time for minority carriers across the base region is a measure of the transient response of a bipolar transistor. The minimum latchup trigger time may be approximated by the sum of the vertical and lateral bipolar transit times.⁴⁰ A typical value of the transit time of a vertical transistor in a 1- μm n -well CMOS technology is several nanoseconds.

6.4.5.2 Value of β in CMOS Lateral Parasitic Bipolar Transistors.

The current gain of a lateral bipolar transistor is determined primarily by the layout spacing between the diffusion outside the well to the well edge, since this is the dimension of the transistor's base width (although the gain is also impacted by the field doping outside of the well and the well depth). Since the base width of the lateral transistor is usually much larger than that of the vertical transistor, the β value is generally an order of magnitude lower (e.g., β ranges from ~ 2 to 4 in CMOS structures whose n^+ to n -well spacings vary from 5 μm down to 2 μm).

The larger base width also means that the base transit time is also much greater. The poor current gain of the lateral bipolar transistor reduces the tendency of the CMOS structure to undergo latchup when dc signals are applied, while the longer base-transit time increases the minimum triggering time for transient-induced latchup. As the

layout spacing shrinks in high-density ULSI, however, the β is increased and the base transit time is reduced.

6.4.6 Circuit and Device Effects that Induce Latchup

For latchup to occur, the emitter-base junctions of both of the parasitic transistors must be forward-biased. This *circuit condition* can be produced in the latchup structure of CMOS circuits in three ways, each of which can be triggered by various physical stimuli. In addition, virtually all latchup failures occur as a result of transient stimuli. To fully describe the various latchup causing mechanisms, we will refer to Fig. 6-18d (the equivalent circuit model), and to Fig. 6-18a (the CMOS inverter cross-section). The three scenarios^{36,38} that lead to latchup are described in the next paragraphs.

6.4.6.1 An external stimulus forward-biases the emitter-base of one transistor, and its collector current then turns on the second transistor. To explain latchup in such cases, let us assume that the externally applied stimulus is a *voltage overshoot at the output node of an p-well CMOS inverter driver circuit*. This is, in fact, the most common cause of latchup.⁵⁴

The sequence of events is as follows: The *source* region of the PMOS device and the substrate contact are both connected to V_{DD} to ensure that the source-substrate junction is never forward-biased (Fig. 6-18a). The *drain* of the PMOS device, however, is connected to the output of the inverter. If this output is *low* (e.g., 0 V), the drain-substrate junction is reverse-biased, and no latchup can occur – that is, both of the structures that could be emitters for the parasitic *pnp* device (i.e., the source and the drain of the PMOS device) are at potentials lower than or equal to that of the base region. If the output state of the inverter is *high*, the drain-substrate bias should then equal 0 V; in this case there should still be no reason for latchup to occur (i.e., we assume that the output of the inverter is designed to reach V_{DD} in the high-output state).

If a voltage overshoot occurs at the output terminal, however, the output node experiences a condition in which V_{DD} is exceeded. (Overshoots and undershoots are both common, especially at input/output [I/O] device nodes, where signals tend to be noisy). If the overshoot causes the voltage at the drain region to exceed V_{DD} by more than about 0.6 V, the p^+-n drain-substrate junction becomes forward-biased. Holes are injected into the *n*-substrate (the base of the *pnp*; see Fig. 6-18a), from the *p*-type drain region of the output node. (This region behaves like a second emitter to transistor Q_1 .) In essence, a triggering current flows through this emitter.

Some of the holes of this triggering current recombine in the base, while the remainder reach the *p*-well (the collector of the *pnp*). The latter represent the *pnp* transistor collector current, which now flows both into the base of Q_2 (I_{B2}) and through R_w to V_{SS} . The fraction of the current that flows through R_w causes a voltage drop across it; this voltage is also impressed across the emitter-base junction of the *nnp* transistor (Fig. 6-18a).

If the voltage drop across R_w reaches 0.6 V, the *npn* device will be turned on. That is, the n^+ source (the emitter of the *npn*) will emit electrons into the *p*-well (the base of the *npn*). Some of these electrons will reach the *n*-substrate and will drift out of the V_{DD} terminal. If enough electron current exists in the *n*-substrate and if sufficient resistance, R_{sub} , exists between the V_{DD} contact and the p^+ source, an IR drop will develop, causing the p^+ source to inject holes into the *n*-substrate. This hole current adds to the initial hole current injected from the positively biased drain region. Thus, the positive-feedback scenario between the *pnp* and the *npn* transistors is triggered, and latchup is rapidly induced. A voltage undershoot when the output node of the inverter is in the low state will have the same net effect, except that the *npn* will be the first transistor to turn on.

Transient overshoot or undershoot voltages are a particular problem at the outputs of CMOS driver circuits, since impedance mismatches at the far ends of transmission lines or printed-circuit-board wiring traces result in reflections that return to the driver output node. It is therefore especially important to utilize stringent latchup-prevention measures at the input and output circuitry of CMOS chips (some combination of guard structures and multiple-well contacts should be used; see section 6.4.8.3).

6.4.6.2 An external stimulus causes current to flow through both bypass resistors, forward-biasing one or both bipolar transistors. Triggering mechanisms that can cause currents to flow in both bypass resistors (R_{sub} and R_w) include avalanche breakdown of the well-substrate junction (J_2 , in Fig. 6-18e), photocurrents due to ionizing radiation, and *n*-well displacement currents (due to the charging or discharging of the large well-to-substrate junction capacitance). As shown in Fig. 6-21, the effects of these mechanisms can be modeled by adding a current source, I_o , and a capacitance, C_{ws} , to the equivalent circuit model of Fig. 6-20b. It is evident that leakage or displacement currents can still flow in both resistors, even if the bipolar parasitic transistors are in cutoff.

Examples of the external stimuli that produce these currents are: (a) voltages across the power-supply terminals that exceed the breakdown voltage of J_2 (and hence cause avalanche breakdown current, I_o); (b) ionizing radiation that causes photogeneration leakage current, I_o ; and (c) external voltage transients (e.g., the voltage step function that occurs when a chip is powered-up) that produce displacement currents in the course of charging and discharging C_{ws} .

If the current in the bypass resistors is large enough, it can turn on both transistors. Typically, however, one of the resistances is larger than the other, and the voltage drop across the larger resistance causes the transistor to which it is connected to be turned on first. The positive-feedback mechanism that induces latchup is then set in motion, and when *Latchup Condition #2* is satisfied (as given by Eq. 6-8), the circuit latches up.

6.4.6.3 Current is shunted through one of the parasitic transistors by some degradation mechanism, and the resulting collector current flows through the bypass resistor of the second transistor, turning it on. The triggering mechanisms that effectively shunt



• • • • •

Abstract

1000

Table 6.1 Latchup Stimuli

I/O node voltage overshoot and undershoot	Type 1
Avalanche of well-substrate junction (extra-high voltage on some node)	Type 2
Photogeneration (ionizing radiation)	Type 2
Displacement current transient (voltage transient during power-up of chip)	Type 2
Inversion under field oxide	Type 3
Punchthrough between n^+ and n -well	Type 3
Hot-electron-induced current	Type 3

be induced (i.e., it is the current drawn from the power supply just before the test structure enters the latched state). Large values of I_{trig} , I_H , and V_H are desirable for reduced latch susceptibility. The efficacy of the processing and circuit-layout schemes designed to reduce the values of R_{sub} and R_w are therefore evaluated by measuring the values of I_{trig} and V_H . *If it is found that V_H is greater than the power-supply voltage, the circuit is said to be latchup-immune.* This assertion is based on *Latchup Condition #3*. Even if latchup is momentarily induced, it will not be sustained, because the power-supply voltage will be less than V_H .

The susceptibility of CMOS circuits to latchup is often determined experimentally by measuring the total current through the $pnpn$ path while overstressing the anode voltage. With the source and n -well contact (Fig. 6-22) maintained at V_{DD} , the isolated voltage on the p^+ region in the n -well is raised above V_{DD} .³⁹ The value of I_{trig} is then measured for the circuit structure being evaluated. (Once a latchup has been triggered, the stressing voltage on the isolated region is returned to V_{DD} .) The value of V_H is measured by lowering V_{DD} after a latchup.

Techniques for studying the transient behavior of latchup have also been described. Such transient testing may provide better characterization of the latchup that occurs in

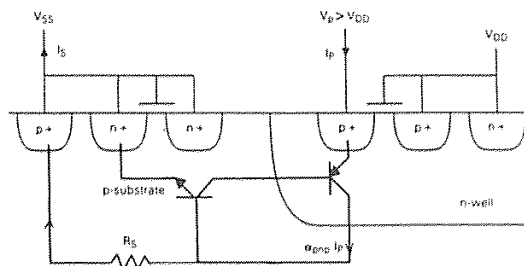


Fig. 6-22 Measurement technique for determining latchup triggering by p^+ overvoltage.

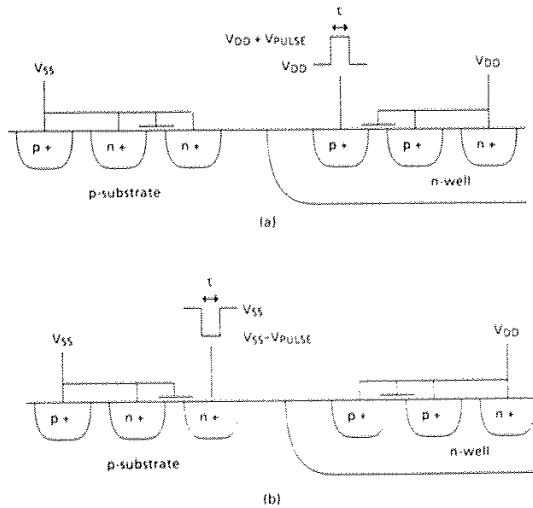


Fig. 6-23 Transient latchup measurement setups: (a) p^+ overvoltage triggering; and (b) n^+ overvoltage triggering.¹¹⁶ (© 1987 IEEE).

actual circuit environments (*Latchup Condition #4*). One test involves the measurement of the *pnpn* diode current when different voltage ramp rates are impressed on the isolated p^+ region of Fig. 6-18a. This allows characterization of the displacement-current-triggering event, described as being one of the causes of Type 2 mechanisms. In another test setup, an overvoltage *pulse* is applied to the drain of the device described in Fig. 6-23. For a given pulse height, the pulse width is increased until latchup is triggered. As the pulse length approaches the bipolar transient response time, the pulse height needed to induce latchup increases. When the length is further reduced, latchup cannot be triggered, regardless of how large a pulse amplitude is applied.

A third transient test for latchup involves pulsing the base of one of the parasitic bipolar transistors.⁴⁰ Transient excitation with a pulse width shorter than the minimum regeneration time causes no latchup.

Another technique for evaluating the latchup hardness of a CMOS circuit design is given by Troutman,³⁶ who defines a differential latchup criterion. The following data need be gathered when this technique is used: the bypass resistance values; small-signal alphas of both parasitic transistors; base-emitter saturation currents for both transistors; and temperature.

6.4.7.1 Modeling Latchup in CMOS Technology. Modeling of latchup has been attempted by a number of different groups. It has turned out to be a difficult undertaking, since the phenomenon involves bipolar transistors with strong injection

effects, with the structures also inherently distributed in nature. As a result, conventional one-dimensional device analysis is very difficult to apply.

6.4.8 Techniques for Reducing or Eliminating Latchup Susceptibility

Circuits are described as being either latchup-free or latchup-immune; the distinction between the two terms is important. *Latchup-free* refers to the ideal situation in which latchup will never occur under any circumstances. *Latchup-immune* refers to circuits that will not exhibit latchup under normal operating conditions, but that could be forced to do so through the application of sufficiently high voltages or the injection of high currents. While as of this writing there is no industrial standard for latchup hardness, a circuit is generally recognized as being latchup-immune if it can withstand an I_{trig} of more than 500 mA at the I/O pins.

Latchup immunity exists if V_H is less than the power-supply voltage per *Latchup Condition #3*. Many reports in the literature describe such a circuit as being *latchup-free*. However, Troutman, points out that even transient switching (unsustained latchup) can cause circuit problems and undesired power dissipation.³⁸ *Thus, he argues that the most effective way to prevent latchup problems is to ensure that the npnp structures remain in the OFF state at all times.*

The approaches to reducing or eliminating susceptibility to latchup can be divided into three categories, as follows:

1. Processing techniques that reduce the current gains of the parasitic transistors to the degree that *Latchup Condition #2* cannot be satisfied.
2. Processing schemes that either reduce the values of R_{sub} and R_w or eliminate the parasitic *npnp* diode structure.
3. Circuit-layout procedures that decouple the parasitic bipolar transistors.

6.4.8.1 Processing Techniques That Reduce Current Gains.

Several techniques have been utilized in an attempt to reduce β of the parasitic bipolar transistors (bipolar spoiling). The first group involves methods that physically separate the emitter and collector regions of the lateral transistor (in which the n^+ regions are kept far from the n -well border). Obviously, for high-density circuitry, merely increasing the spacing of these regions on the wafer surface is not a good solution. Another approach for keeping these regions apart is to use recessed oxides and/or trench structures to increase the distance carriers must travel from the n^+ region to the well. While these techniques do help, by themselves they provide inadequate latchup protection. (These techniques will be described in more detail in the sections dealing with device isolation).

The techniques in the second group attempt to reduce the minority-carrier lifetimes in the base. These include gold doping to produce trapping sites,⁴¹ neutron irradiation to cause structural damage and resultant recombination centers,⁴² and harnessing of the oxygen precipitates formed in the substrate bulk by internal gettering processes for use

as efficient recombination centers for minority carriers.⁴³ The first two of these lead to increased leakage currents, while the third has little effect on lateral transistors with small base widths, since the precipitates do not form in the lightly doped epi. As a result, the techniques in this group have not been widely implemented.* The third type of technique involves the use of a retrograde well, which reduces the vertical bipolar gain by providing a high Gummel number and a retarded E-field in the base.⁴ However, care must be taken to remove injected carriers so that lateral transistor action to the well edge is not increased.³⁹ This approach also requires the availability of a high-energy implanter.

The fourth type involves the use of silicided source/drains^{45,46} to reduce the emitter efficiency of the parasitic bipolar transistors. These techniques have the disadvantage of degrading the gain of the MOS devices.⁴⁷

In summary, the methods described above can help to reduce latchup susceptibility, but provide an insufficient degree of latchup protection. The remaining two categories of latchup-suppression techniques, have been found to be much more effective, and as a consequence, they have been much more widely implemented.

6.4.8.2 Processing Techniques That Reduce R_{sub} and R_w or Eliminate the $pnpn$ Structure. Techniques that reduce the series resistance values include the use of *epitaxial layers on heavily doped substrates* and the use of *retrograde implants*. Techniques that eliminate the $pnpn$ path use either: (1) a combination of *deep trench isolation structures* and *epitaxial layers on heavily doped substrates*; or (2) *silicon-on-insulator (SOI)* substrates.

The use of lightly doped epitaxial layers on heavily doped substrates is effective in reducing latchup susceptibility for two reasons. First, the highly doped layer substantially reduces the value of R_{sub} by placing a low-resistance path for *majority carriers* in the substrate in parallel with the more lightly doped epi region in which the devices are formed. Hence, it very effectively shunts the lateral parasitic bipolar transistor. Second, any *minority carriers* injected into the epi layer that then diffuse into the highly doped substrate are more rapidly recombined there, so that fewer reach the collector of the lateral bipolar device.

Figure 6-24 shows the trigger current and the holding voltage as functions of n^+ to p^+ spacing, d , for various epi thicknesses.⁴⁸ Thin epi can be seen to dramatically improve latchup immunity from either standpoint. For a 12- μm epitaxial layer, the triggering current required for latchup is less than 1 mA when $d = 10 \mu m$. When the epitaxial thickness is reduced to 3 μm , d can be reduced to 5 μm , and I_{trig} is increased to 80 mA. The advantage diminishes as d approaches the epi thickness. However, the minimum epi thickness is limited by outdiffusion of impurity atoms from the heavily

* An approach to creating oxygen precipitates in a thin layer 2.5 μm below the surface by means of ion implantation of oxygen and epi growth has recently been reported.⁴⁴ Since these precipitates are positioned so that they reside in the base region of the lateral pnp device, they are effective in reducing its α by decreasing the minority-carrier lifetime in the base.

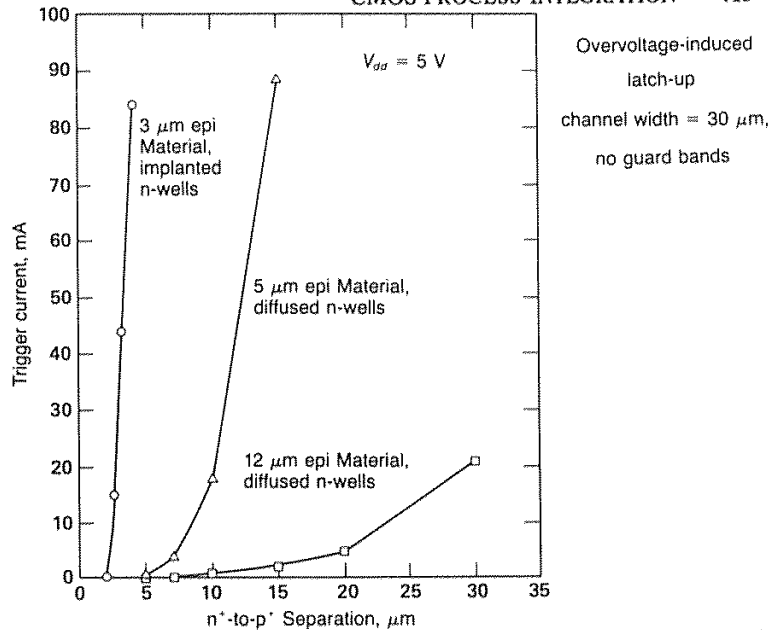


Fig. 6-24 Triggering current versus n^+ -to- p^+ separation in an n -well CMOS structure.⁴⁸ (© 1985 IEEE).

doped substrate during epi deposition and subsequent wafer processing (see section on autodoping in Vol. 1, chap. 5). That is, due to this effect the *effective* epi thickness is thinner than the *initial* epi thickness. (The effective epi thickness is defined as the thickness of the epi layer when the doping concentration is less than some specific number — e.g., $1 \times 10^{18} \text{ cm}^{-3}$.) The fact that n -epi on n^+ produces a sharper epi interface than p -epi on p^+ implies that n -epi should provide better latchup protection.⁴ But as noted in section 6.2, on well technology in CMOS, this benefit is obtained only at the price of accepting the other drawbacks inherent to n -epi on n^+ wafers.

The disadvantages of epi CMOS include increased wafer cost, lower breakdown voltages, and increased leakage currents.⁴⁹ In addition, epi-layer growth may generate defects that will lead to reduced chip yields.⁵⁰

With respect to *retrograde implants*, it has been found that retrograde wells can be effective against latchup for several reasons.⁴ In p -well technology, the retrograde implant is used to form the p -well, with the following advantages being gained:

- The retrograde profile gives a high doping concentration deep in the well, which reduces the gain of the vertical bipolar device.
- The heavier doping near the bottom of the well reduces R_w , thereby helping to decouple the bipolar transistors.

- The retrograde well can minimize thermal processing following epitaxial-layer deposition. Epi layers with thinner *effective* thicknesses may therefore be possible if retrograde wells are combined with epitaxial layers.
- If the combination of the two techniques is used, a process that also reduces *both* R_{sub} and R_w is obtained.

In *n*-well technology, the retrograde implant is a *p*-type implant that is placed beneath the NMOS devices in the substrate.⁵¹ This creates a p^+ layer in the substrate that has a much smaller transition region than a *p*-epi-on- p^+ substrate, thus providing even better latchup protection.

Trench isolation, as described in chapter 2, allows n^+ and p^+ regions to be placed close to one another in CMOS. If the trenches are deeper than the well regions, they provide physical separation among device types. However, when trenches are used without epitaxial layers on heavily doped substrates, their main contribution to latchup immunity is that they force carriers in the lateral transistor base region to travel greater distances to reach the collector. As a result, trench isolation alone does not significantly increase I_{trig} or V_H . The major advantage of using trenches on their own is the increase in latchup response time, which results in a substantial improvement in *transient upset* prevention. When deep trenches are combined with a lightly doped epitaxial layer on a heavily doped substrate, however, latchup-free structures are produced.^{11,96} As minority carriers injected into the substrate attempt to diffuse toward the collector, the deep trench forces them into the heavily doped substrate. There they rapidly recombine, substrate and hence never arrive at the collector region.

Silicon-on-insulator technology is also receiving wide attention, since in addition to providing electrical isolation between the MOS devices, it results in latchup-free CMOS structures. Each MOS device is isolated from neighboring devices by the insulating layer; as a result, the *pnpn* path is no longer present.

6.4.8.3 Circuit-Layout Techniques for Decoupling Parasitic Bipolar Transistors. Two types of structures can be incorporated into the circuit layout that will provide latchup protection: *guard structures*, and *substrate and well contacts*.

Guard structures are heavily doped diffused regions that encircle the well. Troutman states that they are the most effective layout procedure available for providing latchup protection,^{39,54} and this claim is supported by the simulated and experimental evidence of Menozzi et al.⁵² Guard-ring structures can be fabricated either outside the well (in which case the guard surrounds the outer edge of the well region), or inside it (in which case the well is placed between the active device regions and the well border). The guard can be either a *minority-carrier* or a *majority-carrier* structure.

Minority-carrier guard rings have a doping type opposite to that of the region in which they are formed (Fig. 6-25). Normally, these guard rings are placed in the substrate outside of the well edge (e.g., an *n*-type diffusion in the *p*-substrate of an *n*-well CMOS technology). Since the guard is connected to the power supply in such a way that the *pn* junction it forms with the substrate is reverse-biased, this type of guard

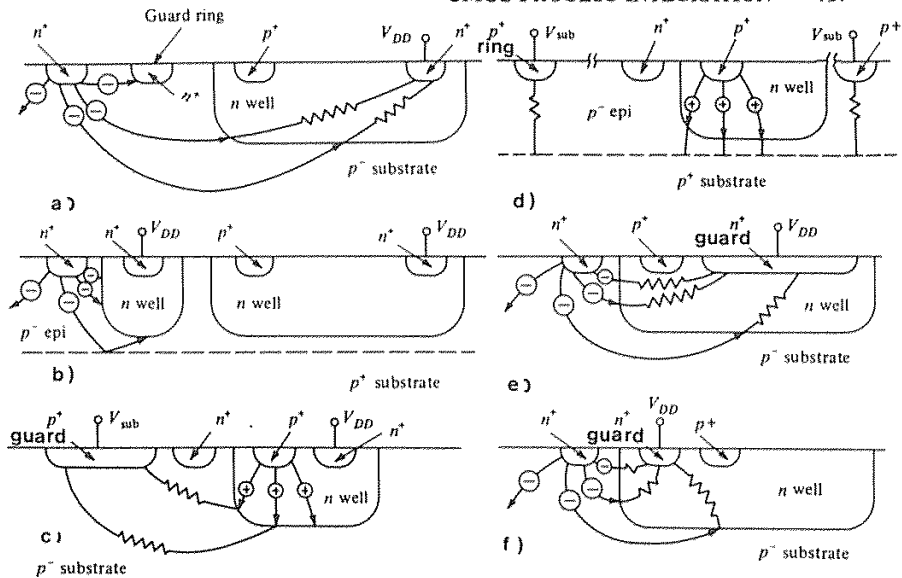


Fig. 6-25 [(a) and (b)] Minority-carrier guard in substrate: (a) n^+ -diffusion guard; (b) Deeply diffused guard in epitaxial CMOS. [(c) and (d)] Majority-carrier guard in substrate: (c) p^+ -diffusion guard for reducing substrate sheet resistance; (d) Contact ring is preferable to p^+ -diffusion guard in epi-CMOS. (e) Majority-carrier guard in well. (e) n^+ -diffusion guard to reduce n -well sheet resistance. (f) n^+ -diffusion guard to steer current away from vertical pnp emitter. ³⁶ From R. R. Troutman, *Latchup in CMOS Technology*, Copyright Kluwer Academic, 1989. Reprinted with permission.

ring will collect any minority carriers in the substrate that diffuse into its depletion region. The carriers are thus prevented from reaching the well, and as a result, the collector current of the lateral bipolar transistor is reduced. The main role of minority-carrier guards in preventing latchup is to provide this reduction in effective current gain.

If a thin epitaxial layer is used together with a deeply diffused guard, the effectiveness of the guard is increased (Fig. 6-25b).³⁶ That is, the path of minority-carrier diffusion is narrowed (by the reflecting boundary where the high and low doped regions in the substrate meet); this forces the carriers closer to the guard ring, where they can be intercepted. In addition, any minority carriers that enter the heavily doped substrate recombine more rapidly there.

Majority-carrier guard rings, on the other hand, are doped with the same type of dopant as the regions in which they are formed. While they can be implemented outside the well, they are usually more effective when placed inside. The function of this type of guard is to provide a shorter (and, hence, a lower-resistance) path for the carriers that constitute the collector current in the well (or substrate). The majority carriers in the collector must drift to the power-supply node through R_w (or R_{sub}) once they enter the

collector region, and the guard rings effectively reduce the value of R_w and R_{sub} by placing smaller resistances in parallel with these series resistances.

The reason that the majority guard is usually placed within the well is that the value of R_w is generally higher than that of R_{sub} . The effective lateral resistance can be quite high in shallow wells, since the source/drain diffusions in the well pinch the lateral current paths. Thus, a guard ring in the well that reduces the distance between the well edge and the well contact can be very effective in reducing the large value of R_w . A way to implement three guard structures with only two diffusions is to place one guard so that it overlaps the well and substrate boundary (Fig. 6-26).

Voltage drops along the power supply bus to which the guard ring is connected should never be allowed to become large enough to initiate a latchup condition. For example, if the bus to which the guard is attached has a sufficiently large resistance that the voltage of the n^+ guard in the n -well drops to 4 V, when the output rises to 5 V, the emitter tied to this higher voltage will turn *ON*, initiating a potential latchup event.

Troutman also emphasizes that a substrate-contact majority-carrier ring should be mandatory for all chips.^{36,39} This would minimize lateral bypass resistance by distributing substrate majority carriers. When used with epitaxial CMOS, such contact rings can reduce lateral bypass resistance to below 1 Ω ; in addition, they are as effective as backside substrate contacts in eliminating latchup. In fact, the need for multiple majority guard rings (which would surround each of the well regions) can be eliminated through the use of a single substrate contact ring (as long as it is used together with wafers that have a thin epi layer on a highly doped substrate).

The major limitation of guard rings is that they decrease overall circuit density. In the input/output circuits of CMOS, however, the MOS transistors must be made quite large in order for sufficient off-chip drive-current capability to be provided. Thus, the area penalty imposed by guard-ring use around these circuits is usually quite small. Because of the noisy signals encountered at the I/O nodes of a chip, the use of guard rings to suppress latchup in such circuitry is widespread.⁵⁴ Each of the output devices in the substrate is typically surrounded by a majority-carrier guard ring (tied to V_{DD} in p -well CMOS, and to V_{SS} in n -well).

Note that n -well CMOS generally provides better latchup immunity than p -well, for two reasons.⁸ First, since the mobility of majority carriers is higher in an n -well than

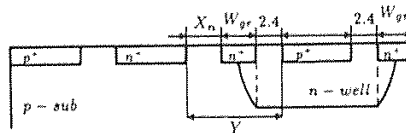


Fig. 6-26 Implementation of two guard structures with only one diffusion by placing one guard so that it overlaps the well-substrate boundary.⁵² (© 1987 IEEE).

in a p -well, R_w in the n -well is lower than in a comparable p -well. Furthermore, R_w is usually much higher than R_{sub} , especially in thin wells that are pinched by source and drain diffusions. Thus, the lower value of R_w in the n -well can be of significant help. Second, because hot-electron-induced substrate currents are much higher in NMOS than PMOS devices, it is better to have the NMOS devices in the substrate. Such currents are more easily collected from here than from the well region, especially if an epitaxial layer on a heavily doped substrate is used.

The second type of circuit-layout technique to be discussed is that of *substrate and well contacts*. The use of guard rings in the well may represent too large an area penalty because in high density circuits multiple well contacts can serve as an alternative approach to improving latchup immunity. However, this approach is not as effective as the use of guard rings.⁵²

The number and placement of power-supply contacts to both the well and the substrate also impact the effectiveness of this approach. Increasing the number of well contacts to V_{DD} reduces latchup susceptibility, since the resistor length of R_w for each FET in the well is reduced when a well contact is in its proximity. The contact spacings should therefore be no more than two squares apart, to ensure that no local high-resistance regions exist. The well contacts should be hard-wired to the power-supply connection (ground or V_{DD}) with a metal stripe. In this way, any injected charge will be shunted to the supply rail through a low-impedance path that will not contribute to the well's already relatively high lateral resistance.

The closer the contacts are to the source regions of the MOS devices, the smaller the potential drop during current flow, leading to increased latchup immunity. The use of *butted source-substrate* and *butted source-well* contacts, in which the n^+ and p^+ regions are contiguous and connected to the same potential has thus become popular.¹⁰¹ While such butted contacts also conserve area, they are limited to FETs that operate with grounded sources.

6.5 CMOS ISOLATION TECHNOLOGY

In CMOS ICs, *like* kinds of devices within a given well must be isolated in the same manner as the devices in either NMOS or PMOS circuits (i.e., through a combination of a thick field oxide and channel-stop doping). However, the isolation requirements of CMOS technology extend beyond those of either PMOS or NMOS alone, in that in CMOS it is also necessary to isolate the p - and n -channel devices from one another. The isolation of p -channel from n -channel devices must satisfy two requirements: (1) any possible leakage currents that could flow between adjacent PMOS and NMOS devices must be suppressed; and (2) the susceptibility of CMOS to latchup must be minimized.

In CMOS structures, the isolation spacing between the n - and p -channel devices is defined as the total of the distances between the edge of the n^+ region of the n -channel device and the edge of the well, and the edge of the p^+ region of the p -channel device and the edge of the well (or, in other words, the n^+ to p^+ spacing – Fig. 6-27).

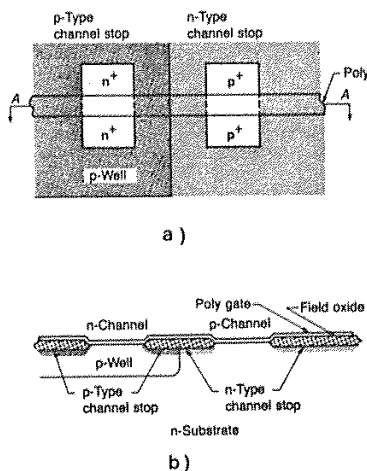


Fig. 6-27 (a) Layout top view of the isolation region between n -type and p -type transistors in CMOS. (b) Cross sectional view of a CMOS inverter showing the channel stops that are designed to prevent the surface under the field oxide from inverting.⁵³ (© 1986 IEEE).

Basic isolation between p - and n -channel devices in CMOS is established by both a reverse-biased well-substrate junction and by the regions under the field oxide (as long as they are not inverted). This isolation, however, may not be perfect. Unwanted current flow can arise due to junction breakdown, the formation of leakage paths between devices, or latchup. The key *leakage paths* involve the *parasitic MOS field devices* created in the field regions between the devices.* Leakage currents can arise if these well-border parasitic field devices conduct prematurely, as a result of surface inversion, or punchthrough below the surface.** As noted, the *latchup* effect in CMOS arises as a

* The parasitic NMOS field transistor at a p -well border consists of the n^+ source, the p -well body, the n -substrate drain, and the polysilicon gate runner over the field oxide, as shown in Fig. 6-27b. Similarly, the parasitic PMOS field transistor at this well border consists of the p^+ source, the n -substrate body, the p -well drain, and the poly runner over the field oxide.

** In Fig. 2-4 of chapter 2 we described how punchthrough can occur between the source and drain of the same device, or between the source/drain regions of neighboring devices of the same channel type. In this case, we refer to the punchthrough between the border of the substrate and source/drain regions in the well.

result of the *parasitic bipolar devices*. Various isolation techniques have therefore been developed to prevent both leakage and latchup. Device and process simulators used to model such CMOS isolation structures are described in chapter 9.

Interest in n - and p -channel isolation techniques is very keen, because such isolation requires much more area than between like types of devices. For example, in early single-well CMOS technologies, the isolation spacing required was typically about three times the diffusion depth of the well.⁸ Thus, for a $4\text{-}\mu\text{m}$ well depth, a minimum n^+ to p^+ spacing of $\sim 12\text{ }\mu\text{m}$ was needed. In early twin-tub CMOS technologies, which used dual-channel stops, a minimum of $3\text{-}9\text{ }\mu\text{m}$ of lateral space was necessary for effective isolation.* In chapter 2, however, we showed that isolation between *like* kinds of MOS devices can be accomplished with isolation spacings of only $1.0\text{-}1.5\text{ }\mu\text{m}$. The large area penalty of p - to n -channel device isolation is one reason why CMOS technologies using conventional isolation methods cannot achieve as high a packing density as NMOS.

This isolation-spacing requirement is due in part to the processes used to fabricate CMOS devices. Wells are typically driven in quite deeply to ensure that enough charge exists below the transistor to prevent vertical punchthrough to the substrate. This results in both a lateral diffusion of the well dopant and a reduction in the surface concentration near the border of the well (Fig. 6-28). The channel stop doping in the

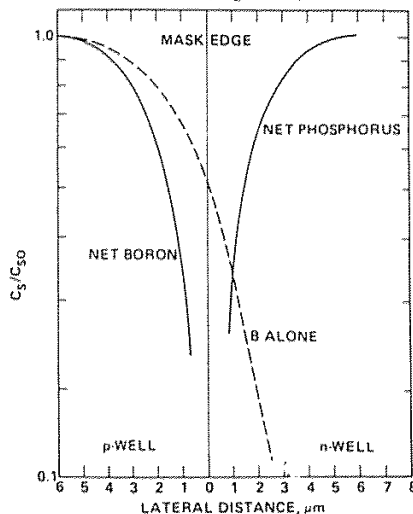


Fig. 6-28 Impurity surface concentration near the border of the two tubs in twin-tub CMOS.¹⁴ (© 1980 IEEE).

* A $3\text{-}\mu\text{m}$ minimum space between n^+ and p^+ regions was reported for a non-trench isolated, twin-well CMOS process. To achieve such tight spacing, high-pressure oxidation was used to grow the field oxide. This was more effective in preventing interdiffusion of the impurities from the two wells than an atmospheric-pressure field-oxidation process would have been.²⁹

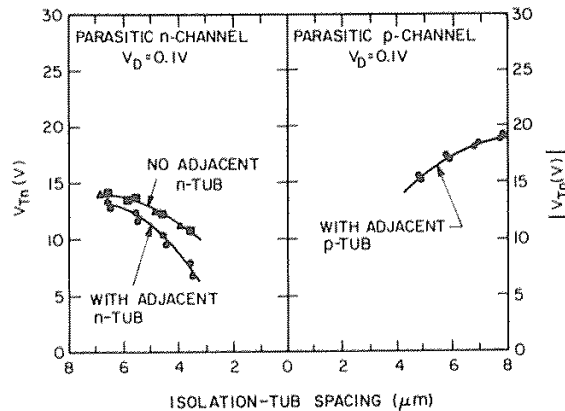


Fig. 6-29 Threshold voltage of n - and p -channel parasitic device as a function of the separation between the transistor edge and the tub border.¹⁴ The upper curve in the left-hand graph shows the parasitic n -channel threshold-voltage reduction near the tub border when no adjacent n -tub is present. When an n -well is present, the threshold voltage is reduced (lower curve), since the interdiffusion of the two types of impurities reduces the net surface concentration of each dopant near the well border.¹⁴ (© 1980 IEEE).

substrate at the edge of the well is also substantially reduced, due to the compensation between the acceptor and donor atoms; this causes a reduction in the threshold voltage of the parasitic MOS field transistors. We will next describe this phenomenon in more detail for the case of twin-tub CMOS.

Figure 6-29 shows the threshold voltage of each type of parasitic device as a function of the separation between the transistor edge and the tub border.¹⁴ The upper curve in the left-hand graph shows the parasitic n -channel threshold-voltage reduction near the tub border when no adjacent n -tub is present. When an n -well is present, the threshold voltage is reduced (lower curve), since the interdiffusion of the two types of impurities (Fig. 6-28) reduces the net surface concentration of each dopant near the well border. Unless the net dopant-reduction effects are somehow counteracted, the spacing between adjacent n^+ and p^+ devices must be kept quite large (e.g., $>10\ \mu\text{m}$) to prevent inversion beneath the field oxide.

If smaller spacings between n - and p -channel devices are to be possible, the channel-stop doping concentration must somehow be increased, particularly in the substrate regions in n -well CMOS. During field oxidation, boron segregates into the oxide, while phosphorus piles up at the silicon surface. As a result, in an n -well process a separate p -type channel-stop implant must be added to increase the surface concentration of the lightly doped p -substrate. Without such an implant, inversion between the n -channel

devices is likely to occur. In *p*-well technologies the well itself is an adequate *n*-type channel stop, because heavier boron doping exists in the *p*-well, and the concentration of phosphorus is increased at the surface of the *n*-substrate during field oxide growth.

In twin-tub processes, two channel stops are needed to reduce the isolation distance as much as possible. Figure 6-30, which is based on a 2-D device model,⁵³ shows the impurity contours in the isolation region of a CMOS twin-tub process (with retrograde wells).

The AT&T Twin-Tub V technology,⁵⁵ is an example of an advanced CMOS process uses a LOCOS-based approach to isolate like devices. The starting material is a *p*-epi layer on a *p*⁺ substrate. The technology is implemented with a single-mask, self-aligned twin-well process that uses two channel-stop implants (as well as two separate well implants).

First, phosphorus and arsenic are sequentially implanted into the *n*-well areas, forming a high/low doping profile (channel-stop and well; Fig. 6-31). Since arsenic diffuses much more slowly than phosphorus, it will remain near the surface of the well during the drive-in step. This high arsenic concentration provides an effective channel stop for *p*-channel devices and also protects against punchthrough.

Next, a relatively thick masking oxide is selectively grown over the *n*-well region, and the *p*-well implant (boron) is then performed. Because the implant is blocked by the oxide that covers the *n*-wells, it enters the silicon only in the *p*-well regions (a self-aligned implant step). Both well regions are then driven in (with the thick oxide over the *n*-wells being retained), and a second boron implant (which will serve as the *p*-well channel stop) is carried out. This implant is kept shallow because a high pressure LOCOS process is subsequently used to grow the field oxide, minimizing the lateral and vertical impurity-profile spreading. Once again, the implant is self-aligned to the *p*-well regions by the presence of the oxide on the *n*-wells (Fig. 6-31). The masking oxide is then removed, and a nitride masking layer is deposited and patterned to cover the active areas. Finally, the field oxide is formed.

This process makes it possible to achieve adequate isolation and a reduced tendency

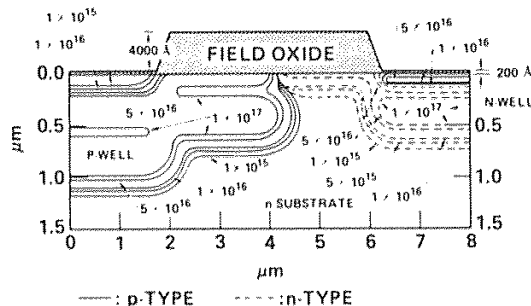


Fig. 6-30 2-D net-impurity contours in the isolation region of a retrograde-well CMOS structure.⁵³ (© 1986 IEEE).

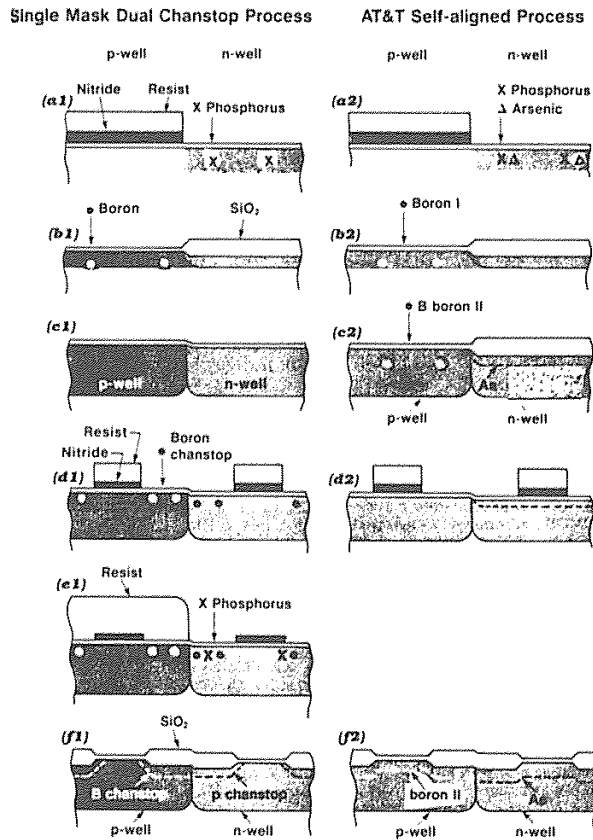


Fig. 6-31 (a1) - (f1) Twin-tub dual-field implant process.¹¹⁹ (a2) - (f2) Self-aligned twin-tub and field-implant process using high-pressure field oxidations.⁶²

toward latchup with smaller device separations than are possible with an atmospheric-pressure oxide-growth process. Nevertheless, a $7\text{-}\mu\text{m}$ spacing must be used between the n^+ and p^+ regions in the Twin-Tub V structure to provide a V_H greater than the power-supply voltage (5 V). Since this spacing still represents a significant layout-density limitation, alternatives to the LOCOS-only process are being vigorously pursued.

In another, more recent study using n -well technology, high-field channel-stop implant doses for the substrate (yielding a peak concentration of 5×10^{16} boron

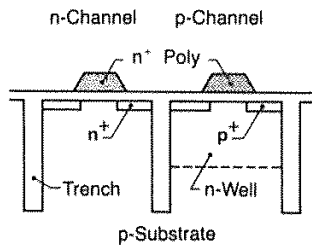


Fig. 6-32 Schematic of an n -well CMOS with trench isolation.⁴ (© 1986 IEEE).

atoms/cm³) allowed n^+ to p^+ spacings of $2.4\ \mu\text{m}$ with conventional wells, and $1.8\ \mu\text{m}$ spacings with retrograde wells.¹⁰⁶

6.5.1 Trench Isolation for CMOS

The two major alternative CMOS-isolation approaches are *trench isolation*⁵⁸ and *selective epitaxial growth* (SEG) isolation. Trench-formation technology, discussed in chapter 2, is applied to CMOS to attack the problems of latchup and punchthrough. Its main advantage is that latchup can be completely eliminated if a process employing a thin epitaxial layer on a heavily doped substrate is used, and if the trench is allowed to penetrate to the heavily doped region.¹¹ Trenches of micron and submicron widths have been used to reduce p^+ to n^+ spacings to $2\text{--}2.5\ \mu\text{m}$.⁵⁷ In addition, when trenches deeper than the well depth are used, they replace the reverse-biased pn junction as the isolation structure at the well sidewalls (Fig. 6-32).

As of 1988, active devices can not yet be set against the trench sidewall, due to channel-inversion problems associated with the vertical trench sidewall (Fig. 6-33).⁵⁶ The sidewall inversion is caused by the horizontal parasitic MOS device, with the well acting as the gate electrode and the trench dielectric acting as the MOS gate oxide (with a thickness equal to the trench width). The voltage across this parasitic device is 5 V

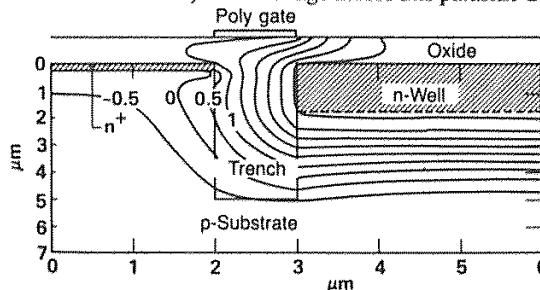


Fig. 6-33 The corresponding two-dimensional potential contours showing sidewall inversion; the simulation is done at $Q_b = 5 \times 10^{10}\ \text{cm}^{-2}$, $N_A = 6 \times 10^{14}\ \text{cm}^{-3}$, $V_{GD} = 3\ \text{V}$, $V_{GS} = 0\ \text{V}$, $V_{SB} = -1\ \text{V}$.¹²⁰ (© 1983 IEEE).

under normal CMOS operating conditions. This gate voltage and the narrow (e.g., 1- μm) trench width can easily cause inversion along the sidewall, outside the well but facing it.

There are two reasons why p -wells exhibit more severe trench-sidewall inversion problems. First, some of the needed boron doping segregates into the trench oxide during the thermal oxidation step. The second reason is the presence of the fixed charge, which is normally positive.

Once sidewall inversion occurs, n -channel devices with regions butted to the same sidewall become electrically connected by a path along the sidewall. Two obvious solutions are to leave a separation between the n^+ and the sidewall or to increase the trench width, with the end result of either approach being an increase in the isolation spacing. Because of this limitation, the minimum n^+ to p^+ isolation distance in trench isolation is 2-2.5 μm (1988).

Trench isolation has several other disadvantages, as follows:

- Fabrication is much more complicated than in LOCOS, making this an expensive alternative approach.
- For adequate filling and planarization, only one trench width can be utilized (unless SEG refill of the trenches is used).
- Another type of isolation is also needed (usually LOCOS) for most inactive parts of a chip. If a trench deeper than the well depth were used to isolate each device within the well region, each device would be isolated from the substrate. To keep the device bodies from floating, it would be necessary to provide each one with its own well contact, rather than using a single contact for the entire well. Chip area would thus be wasted, defeating the purpose for which the technology was adopted.

6.5.2 Isolation by Selective Epitaxial Growth for CMOS

Various schemes using SEG have been explored as CMOS isolation alternatives (see chap. 2 for more details on SEG technology). One of the first, *selective-etch-and-refill-with epi* (SEREPI), uses an epi refill of recesses in silicon, with the sidewalls passivated prior to refilling. In one approach,⁵⁹ the well regions were anisotropically etched to a 5- μm depth, and a 200-nm thermal oxide was grown on the recess surfaces. This oxide was removed from the bottom of the recess with an anisotropic dry etch, leaving it on the sidewalls. Buried layers were formed by arsenic implantation into the bottoms of the wells, and then epitaxial silicon was selectively grown in the wells. By controlling the growth rate, it was possible to refill the wells so that the surfaces were level with the outside substrate.

The SEREPI structure offers the advantages of minimum isolation area and also the flexibility of dopant-profile control in the epitaxial silicon in the wells. A similar process, reported by Kasai et al.,⁶⁰ uses a sidewall layer consisting of a 250-nm-thick composite film of oxide and nitride (see Fig. 2-34b, chap. 2). After these layers have been removed anisotropically through dry etching, a sacrificial oxide is grown on the

bottom of the trenches and is then wet-etched away to remove any dry etching damage. Finally, an SEG step is performed.

While it is possible to make the n^+ to p^+ isolation distance as small as $0.25\text{ }\mu\text{m}$, with either approach, some problems exist. First, LOCOS must still be used to isolate devices from one another within the same well. Second, unless a deep boron implant is added to the n -channel side of the isolation, sidewall leakage or inversion can occur. Finally, the $0.25\text{-}\mu\text{m}$ distance between p^+ and n^+ regions can become a limitation for masking of opposite-type implants if n^+ and p^+ junctions are to be placed on opposite sides of the $0.25\text{-}\mu\text{m}$ -thick dielectric.

Another SEG-based approach for isolating CMOS has been described by Manoliu and Borland,¹⁵ and by Stivers.⁶¹ In this approach, windows are anisotropically etched into a $1\text{-}2\text{-}\mu\text{m}$ -thick surface-layer oxide (see Fig. 2-43, chap. 2). Separate implants are used to form p^+ layers in those recesses designated to be p -wells, and n^+ layers in those that are to be n -wells. The oxide recesses are then refilled, using an SEG step in which sidewall inversion in the p -well is suppressed by tailoring the doping profile of the boron doping (i.e., this region is doped with $1 \times 10^{17}/\text{cm}^3$ in the active area of the n -channel devices). As a result, no subthreshold "kink" is observed in the turn-off current (such kinks in the subthreshold current curve are due to leakage from source to drain along the sidewall of the isolation structure).

The two buried layers increase the latch-up immunity of the CMOS devices considerably. Furthermore, this approach eliminates the need for an additional LOCOS isolation. Although this technique was only tested for minimum spacings of $2\text{ }\mu\text{m}$, it is estimated that it will provide adequate isolation for spacings down to $1\text{ }\mu\text{m}$. Finally, good planarity and gate oxide integrity were observed.

The technique of *retrograde wells* has also been investigated for implementing reduced isolation spacing and increased latchup immunity in CMOS (as well as for enhancing other device characteristics).^{16,19} In this approach, the wells are formed by means of a high-energy implant, which is performed following formation of the field oxide. Because the peak of the implant is well below the silicon surface, the impurity concentration in the well decreases as it approaches the surface (hence the name "retrograde well").

It is easier to implement p -well CMOS with retrograde wells, since boron has a higher implant range in silicon than does phosphorus, and boron implants thus require a lower ion-implantation energy in order to achieve a similar depth. Because the well is formed after the field oxidation, retrograded p -wells can have higher NMOS field-threshold voltages compared to those exhibited in conventionally-formed wells. (In the latter, some of the boron segregates into the oxide during field oxidation, leaving less boron present at the silicon surface. This reduces the field-region threshold voltage.) In addition, the lateral diffusion of boron that occurs when the channel-stop implant is done prior to field oxidation is eliminated.

There are several drawbacks to the retrograde well approach. First, it produces devices with a high junction capacitance (C_j) and a high body-effect coefficient (γ) due to the relatively high doping concentration below the surface in the wells. These effects decrease the speed of circuits manufactured with such devices. Second, this approach

may require the use of implanters with accelerating voltages greater than 400 keV. While such machines are available, their low beam current makes them incompatible with high-volume manufacturing needs. Note that doubly ionized boron can be implanted with an accelerating voltage of 200 keV to obtain the same result, and that such a voltage is within the range of ordinary implanters. However, the beam current in this case remains low (see Vol. 1, chap. 9.)

6.6 CMOS PROCESS SEQUENCES

Since there are so many options available for designing a CMOS process flow, no "standard" approach has been adopted. The process flows presented here are merely illustrations of the general sequence of process steps and of the types (and number) of masking layers used. We will first present examples of simple *p*-well and *n*-well single-level-metal CMOS process flows. Such process sequences might be used in the fabrication of CMOS circuits with a minimum feature size of 1.25-2.0 μm . We will then describe a twin-well, double-level-metal process that uses LDD structures. With additional enhancements such as trench or SEG isolation, the filling of contact holes and vias with CVD metal, and novel interlevel dielectric planarization methods, such a process would likely be used in the manufacture of CMOS circuits with submicron dimensions.

6.6.1 Basic *p*- and *n*-Well CMOS Process Sequences

A basic single-well CMOS process can be implemented in either *p*-well or *n*-well technology using eight masking levels. Figure 6-34 shows seven of the eight masking levels of a basic *p*-well process. Figure 6-35 illustrates the *front end* masking levels of an *n*-well process. (The *back end* steps of a process sequence are those that begin with the contact masking step. Thus, the term *front end processing*, refers to all of the steps up to that point.) We will describe an example *n*-well CMOS process in more detail.

The NMOS devices in the *n*-well technology are formed in the lightly doped *p*-substrate ($\leq 1 \times 10^{15}/\text{cm}^3$), while the PMOS devices are formed in the more heavily doped *n*-well ($\sim 1 \times 10^{16}/\text{cm}^3$). The starting material is either a lightly doped $\langle 100 \rangle$ *p*-type wafer or a heavily doped $\langle 100 \rangle$ *p*⁺ wafer with a thin (5-10- μm thick), lightly doped *p*-type epitaxial layer at the surface. A process for enhancing the gettering capabilities of the wafer may be employed before feature formation on the wafer surface is begun (see Vol. 1, chap. 2).

The *n*-well regions are the initial features formed on the starting material. First, a thermal oxide is grown and a CVD nitride film deposited. Then *Mask #1* is used to pattern windows in these layers, through which phosphorus for the *n*-well is implanted. Since the implantation process is unable to place the phosphorus ions deeply enough into the silicon, these impurities must be driven in to the appropriate depth during

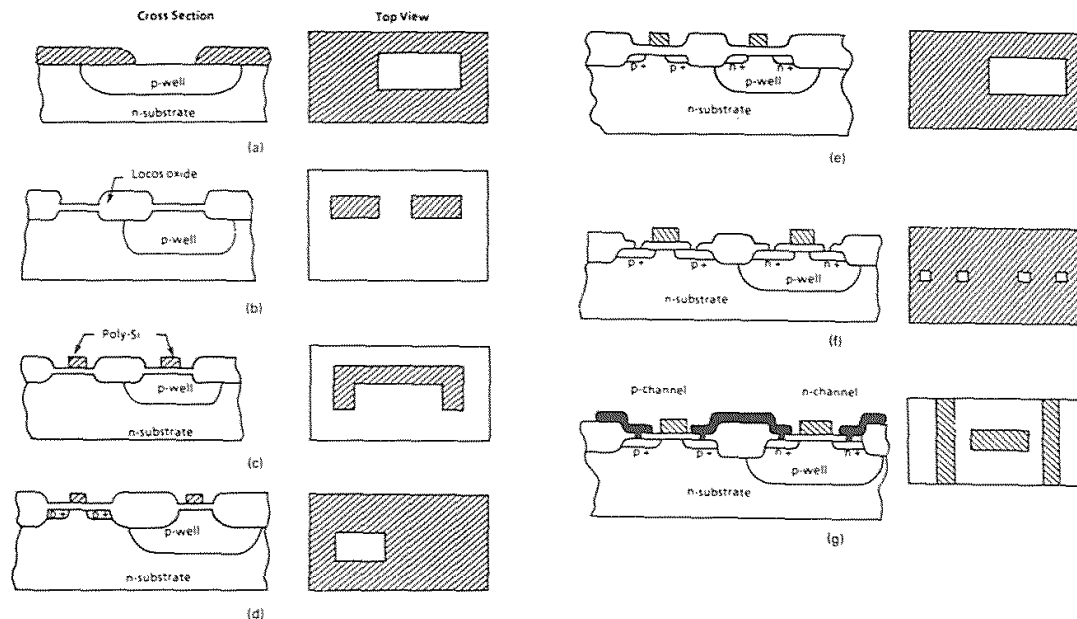


Fig. 6-34 Cross sections and top views of a typical p -well CMOS process at all mask levels: (a) p -well mask; (b) active-area mask; (c) poly-gate mask; (d) p^+ mask; (e) n^+ mask; (f) contact mask; (g) metal mask.¹ From J. Y. Chen, *CMOS Devices and Technology for VLSI*. Copyright Prentice-Hall, 1989. Reprinted with permission.

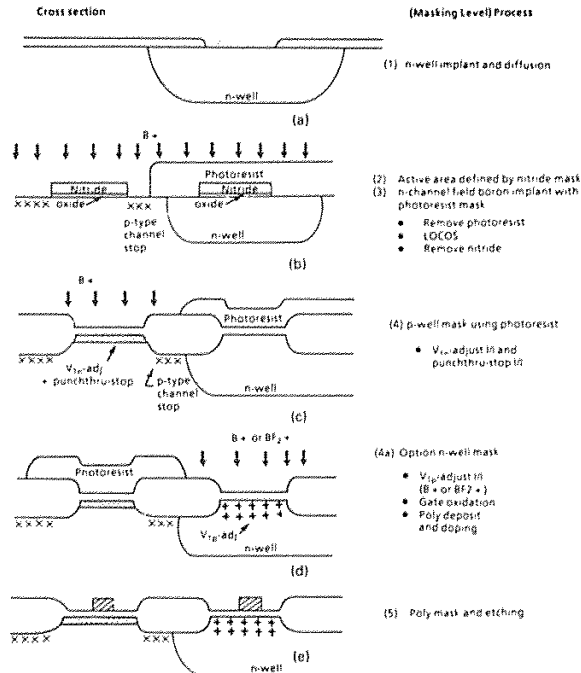


Fig. 6-35 Cross sections and masking levels and associated processes of an n -well CMOS front-end process.¹ From J. Y. Chen, *CMOS Devices and Technology for VLSI*. Copyright Prentice-Hall, 1989. Reprinted with permission.

subsequent high temperature cycles. An oxide is also grown on the n -well regions during the drive-in step. At the conclusion of the drive-in processes the surface concentration in the well is $\sim 1 \times 10^{16}/\text{cm}^3$, and the impurity concentration gradient within the well is also rather small. Note that the redistribution of the well dopant occurs laterally as well as vertically.

Next, a boron threshold-adjust implant is carried out. There is no resist mask for this step, as the thin oxide or oxide/nitride layer covering the silicon wafer surface protects it from contamination. As was described in section 6.3.1, this single implant can provide a correct V_T adjustment for both the NMOS and the PMOS devices.

The surface is then stripped of its oxide and nitride/oxide layers, and a new pad-oxide/nitride layer (needed for LOCOS) is formed. *Mask #2* is then used to pattern this layer to define the active device and field regions. A boron channel-stop implant is performed for the p -substrate field regions. Although no separate mask is used, the boron implanted into the *well field regions* is not of sufficient concentration to significantly alter the n -concentration there. In addition, when the field oxide is grown the phosphorus piles up beneath it in the well regions (see Vol. 1, chap. 7), while some of the boron implanted during the channel-stop implant segregates out into the field

oxide. Hence, the surface concentration of phosphorus in the well remains high enough that a separate channel-stop implant is usually not needed for the well region.

The field oxide is then grown, after which the nitride/oxide layer is removed from the active device regions (see chaps. 2 and 5). Next, a gate oxide is grown, in the same manner as that described in chapter 5 for NMOS devices. (Note that a *sacrificial pre-gate oxide* is frequently grown and stripped prior to the growth of the actual gate oxide.)

The deposition of polysilicon by CVD for the gate layer is then carried out. This layer is subsequently doped with phosphorus to form an n^+ polysilicon gate material. The resistivity of the polysilicon should be as small as possible, since this layer also serves as an interconnect structure. (In more advanced processes, even lower resistivities are achieved by forming a silicide layer on top of the polysilicon layer.) *Mask #3* is used to pattern the polysilicon.

Masks #4 and *#5* are then used to selectively implant the source/drain regions of the PMOS and NMOS devices, respectively. The polysilicon protects the channel region under the gate from being implanted. Arsenic is preferable for the n^+ regions so that shallow junctions and minimum lateral diffusion under the gate can be obtained. (It is typically implanted to a dose of $3\text{-}6 \times 10^{15} \text{ cm}^{-2}$, and with an energy of 40-60 keV.) Boron is often implanted as BF_2^+ (for shallow junction formation) at doses of $1\text{-}5 \times 10^{15} \text{ cm}^{-2}$ and energies of 30-50 keV. Typical values of source/drain sheet resistance should be below $30 \text{ } \Omega/\text{sq}$. Ohmic contacts to the n -well and p -substrate are formed simultaneously with the implants that create the NMOS and PMOS source/drain regions, respectively. These implants are then annealed with a short thermal process at a moderate temperature (e.g., 900-1000°C). The gate oxide that covers the source and drain regions during the implant is usually later stripped and regrown, since it has been damaged by the heavy implants and may have been contaminated by the RIE step used to pattern the poly.

A CVD doped oxide (with the dopants being either phosphorus, or boron + phosphorus) is deposited to a thickness of 50-100 nm, to serve both as a dielectric between the polysilicon and metal layers, and as a gettering layer. This layer is flowed (see Vol. 1, chap. 6) to improve the wafer-surface topography with respect to metal step coverage.

Mask #6 is used to open contact windows in the CVD oxide so that connections can be made between the metal layer and the silicon and polysilicon. Following a *reflow step*, an aluminum (or aluminum-silicon) layer is deposited and patterned (using *Mask #7*). The wafers are subjected to a final anneal step, which is followed by the deposition of the passivation layer. *Mask #8* is then used to open windows in this layer.

An important aspect of this simple CMOS process is that it uses only one more mask than the E-D NMOS process described in chapter 5. The three extra masks in the CMOS sequence (two masks for the source/drain implants, and one for the well mask) replace the depletion-mode implant mask and the buried-contact mask of the E-D NMOS process. Thus, the process complexity of a modern simple CMOS process is not much greater than that of advanced NMOS.

6.6.2.1 Starting Material. The starting material in an advanced twin-well process is typically a $\langle 100 \rangle$ -orientation, heavily doped substrate on which a thin (5-10- μm thick), lightly doped epitaxial layer has been grown. Although n -epi-on- n^+ substrates have some advantages for a few special applications (e.g., for building CMOS SRAMs; see chap. 8), p -epi-on- p^+ substrates are the more common choice, because they are less sensitive to process-induced defects (see section 6.2).

[illegible]

Dell Ex. 1008B
Page 166

In this method, a single mask is used to pattern a nitride/oxide film that has been formed on the bare silicon surface (*Mask #1*). The openings in the film become the *n*-well regions, and phosphorus is then implanted into them (e.g., at 80 keV). Next, a thermal oxide is grown on these regions, to a thickness that is sufficient to block the boron implant used to form the *p*-wells. After stripping of the oxide/nitride layer, the silicon in the *p*-well regions is exposed, while the *n*-well regions are covered with the implant-blocking oxide. Thus, during the *p*-well implant (e.g., at 50 keV), the boron penetrates the silicon only in the desired areas.

Next the wells are driven in (e.g., at 1100°C for 500 min). At the conclusion of the drive-in steps, the concentration in the wells for a 0.8- μm -CMOS process could be $\sim 1 \times 10^{16}/\text{cm}^3$ for the *p*-well, and $\sim 3 \times 10^{16}/\text{cm}^3$ for the *n*-well. The *n*-well might have a higher doping concentration, to improve the punchthrough performance of the PMOS devices and to eliminate the need for a separate channel-stop step for the *n*-well. A higher concentration in both wells would still produce devices with relatively low capacitances at the bottoms of the source/drain-to-well junctions.

The channel-stop implant procedure is also usually included in the part of the processing sequence in which the wells are formed. In one reported twin-well process,²⁶ only a *p*-well channel-stop implant (boron) is used, because the doping in the *n*-well is high enough ($3 \times 10^{16} \text{ cm}^{-3}$) that a separate channel stop implant is not required. In this case, an unmasked boron implant is done following both well implants, but prior to field-oxide growth.

In a second approach,⁵⁵ an additional mask can be used to provide both *n*-well and *p*-well channel stops (Fig. 6-31a). Note that in this procedure, the boron channel stop is implanted into both the *n*-well and *p*-well field regions; the phosphorus channel stop must be increased to compensate for the presence of the boron in the *n*-well field regions. The disadvantages of this method include the additional alignment step between the channel stops and the well masks, the interdiffusion that occurs during the oxidation step, and the asymmetry of the doping profiles. (The latter is due to the increased phosphorus concentration that must overcompensate the nonselective boron channel stop).

A maskless variation of this method has been developed to overcome the above drawbacks (Fig. 6-31b).^{55,64} In this sequence, both arsenic and phosphorus are implanted into the *n*-well regions. This places both the dopants that form the well and the dopants that form the channel stop into the *n*-well regions prior to implanting of the *p*-wells. An oxide is then selectively grown on the *n*-well regions, and the boron dopant for the *p*-wells is implanted. After the *n*-wells have been driven in, the oxide over the *n*-wells is retained, and a second boron implant is carried out. This implant serves as both a channel stop in the *p*-well *field* regions and a punchthrough-prevention implant in the *active* regions of the *p*-well.

Various approaches have been proposed to prevent excessive redistribution of the boron channel stop implant during field-oxide growth. These include:

1. Co-implanting of Ge with the boron, which has been found to reduce the boron diffusion constant.⁶⁵
2. Implanting of the field regions with Cl to increase the field-oxide growth rate

(thus reducing the amount of time the wafers must be exposed to the high temperature).⁶⁶

3. The use of high-pressure oxidation to reduce the thermal cycle that must be employed to grow the field oxide.^{12,26,55} This also reduces the boron channel-stop dopant loss to the growing field oxide, as well as the interdiffusion of the wells (As a result, smaller isolation spacings between *n*- and *p*-channel devices can be used).

6.6.2.3 Definition of Active and Field Regions. In a conventional advanced CMOS process, a standard LOCOS process (described in detail in chaps. 2 and 5) is used to define the active and field regions (*Mask #2*) and to grow a thick thermal oxide over the field regions. Enhanced LOCOS approaches, which overcome some of the limitations of conventional LOCOS, can be used as alternative processes to form isolation structures in the field regions. Some that have been reported as having been integrated into submicron CMOS processes include the SILO process⁶⁷ and a poly-buffered LOCOS process²⁶ (see chap. 2 for details on both of these). A high-pressure oxidation process for growing the field oxide, was reported to have been used in both of these processes.

6.6.2.4 Gate-Oxide Growth and Threshold-Voltage Adjustment. The gate-oxide growth process is essentially the same as that used in the NMOS process described in chapter 5. The threshold-voltage implant process can be accomplished with either a single implant step, using boron (see section 6.3.1), or two separate implants (in which case additional masking steps are required). A separate selective punchthrough-prevention implant for the NMOS devices is usually needed, especially for submicron devices; this step also requires the use of an extra mask. The V_T -adjust and punchthrough-prevention implants can also be done either before or after the gate-oxidation step. To keep the boron V_T implant shallow (i.e., to improve the PMOS punchthrough characteristics, as described in section 6.3.2), implanting is sometimes done through the gate oxide. To protect the gate oxide from contamination (as well as to keep the implant shallow), part of the polysilicon can be deposited prior to the implant step,^{26,67} with the remainder deposited afterward.

Alternative gate-oxide materials are being studied as replacements for SiO_2 . Gate oxides of less than 10 nm will not be practical unless the gate voltage is also reduced, for two reasons. First, in order to prevent tunneling, the minimum gate oxide thickness must be $2V_{DD}/E_{ox}(\text{max})$, where $E_{ox}(\text{max})$ is 6 MV/cm. For $V_{DD} = 3$ V, the minimum thickness is thus 10 nm. The second reason has to do with the reliability and burn-in techniques used in accelerated lifetime testing. It becomes very difficult to screen out weak devices from good ones when the breakdown-voltage criterion defining a weak oxide falls within the range of the distribution exhibited by the good devices. This situation is encountered at oxide thicknesses below 10 nm.⁶⁸

If a material with a larger dielectric constant than that of SiO_2 (3.9) could be used, the gate dielectric could be thicker, while the same capacitance per unit area could be maintained. Extensive work has been conducted on such materials, including nitrided

oxides, thermal nitrides (dielectric constant = 8), and tantalum pentoxide, Ta_2O_5 (22). Rapid thermal processing has been explored as a means for forming the nitrided oxides, since the high temperatures needed for their formation can be produced by rapidly heating and cooling the wafers.

The Ta_2O_5 films have been prepared through thermal oxidation of a tantalum layer,⁶⁹ by reactive sputtering,⁷⁰ and CVD deposition.⁷¹ Several reports indicate that these films exhibit excellent characteristics, and thus show promise for potential applications in advanced integrated circuits.^{72,75}

In addition, a novel technique for forming ultrathin (15 nm), low-defect SiO_2 gate layers has been reported.⁷⁶ In this approach, a 5-nm thermal oxide is first grown. A 5-nm CVD oxide is then deposited, and finally another 5-nm oxide is thermally grown, *under* the CVD layer (Fig. 6-37). The layer is able to grow beneath the CVD oxide because the oxidizing species diffuse through the first two layers to the Si- SiO_2 interface (see Vol. 1, chap. 7). During the second thermal oxidation, the top CVD oxide is also densified. This combination layer exhibits a low defect density because the growth-induced micropores (which are the major factor contributing to the defect density in gate dielectrics), are misaligned. If they were to extend completely through the oxide film, these micropores would be potential paths for rapid diffusional mass transport, as well as for current leakage.

6.6.2.5 Polysilicon Deposition and Patterning. The process of forming the n^+ polysilicon gate structure involves the deposition of a polysilicon layer and patterning with *Mask #3*. The steps followed are essentially the same as those in the basic CMOS process. In most advanced CMOS processes the polysilicon film is overlaid with a refractory metal silicide to form a *polycide* structure (see Vol. 1, chap. 11). In some cases, a *salicide* (self-aligned silicided gate and source/drain regions – see Vol. 1, chap. 11) process is employed to reduce the parasitic resistance of the source and drain regions, as well as that of the gate material. In most such cases, the salicide

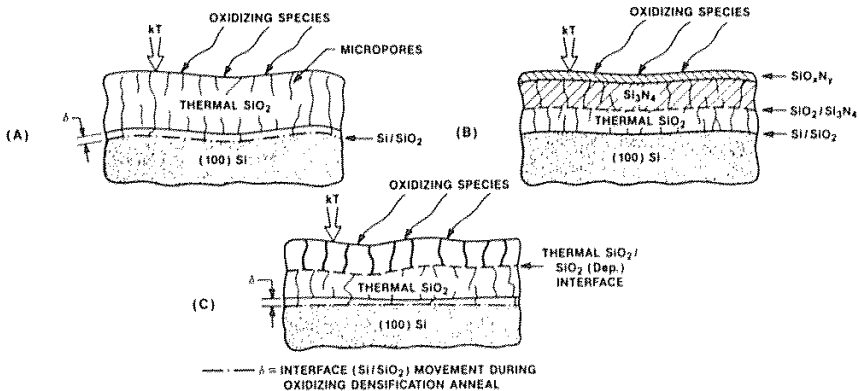


Fig. 6-37 Effects of oxidizing densification anneal: (a) SiO_2 films; (b) SiO_2 - Si_3N_4 dual dielectric; (c) thermal SiO_2 - deposited SiO_2 multilayered stacked films.⁷⁶ (© 1988 IEEE).

formation is carried out after the source and drain have been implanted.

The polysilicon can be doped either after deposition by diffusion or ion implantation, or *in situ* during deposition. The former approaches were used in most MOS processing prior to submicron device generations. For submicron devices, in which a reduced thermal budget becomes more imperative, *in situ* doping becomes more attractive, (since the long, high-temperature drive-in step can be eliminated).¹⁰³ The redistribution of channel and source/drain implants is thus minimized, allowing improved control of threshold voltage, punchthrough, and L_{eff} .

As was described in sections 6.3.3 and 6.3.4, alternative materials to n^+ polysilicon (or to polycides with n^+ polysilicon as the underlayer) have been investigated as gate materials (e.g., p^+ and n^+ polysilicon, or refractory metal gates). Each alternative process introduces some changes into the process flow; nevertheless it is predicted that as device dimensions shrink, such measures will need to be used to maintain adequate device performance.

Etching of the gate structure can be a difficult step, especially when gate lengths are in the submicron range. First, the dimensions must be accurately and uniformly produced across the wafer, as I_D is strongly dependent on the gate dimensions. Second, the sidewalls must be vertical in order to reduce asymmetric I_D characteristics in devices built with LDD structures (see section 5.6.5.2). Third, the formation of vertical sidewall gate structures implies the need for a completely anisotropic etch step (see Vol. 1, chaps. 15 and 16). In turn, such a step requires that the etch process be highly selective against etching the thin underlying gate oxide layer.

6.6.2.6 Formation of Source/Drain Regions. In advanced-CMOS processes, the gate lengths are short enough that LDD structures must be used to minimize hot-electron effects, especially in NMOS devices. Therefore, the procedures outlined in section 5.6.5 are used to fabricate such LDD structures. If these structures are used only for the NMOS transistors, *Masks #4* and *#5* are used to allow the sources and drains of the two transistor types to be selectively implanted.* Note that if LDD structures are also used for the PMOS devices, two additional masking layers may be needed (6-38a).

A removable-spacer LDD process for both NMOS and PMOS devices has been reported (Fig. 6-38b) that does not require the use of any other masks than the two needed to selectively form the sources and drains of the two transistor types.^{67,73} In the removable spacer process, the heavily doped source/drain implant is performed first, with the spacers in place. After the spacers have been removed, the implant that forms the lightly doped drain regions is carried out (Fig. 6-38b). This process can be used to provide LDD structures for one or both transistor types, as desired. Polysilicon is used as the material of the removable spacers in one approach,^{67,99} while a low-temperature

* A procedure that requires only one masking step for creating both types of source and drain regions has also been reported. In this approach, the heavy boron implant is carried out nonselectively. The mask is then used to cover the PMOS device active regions. An n -type implant heavy enough to overcompensate for the boron implant in the active regions of the n -channel devices is used to form the sources and drains of the NMOS transistors.

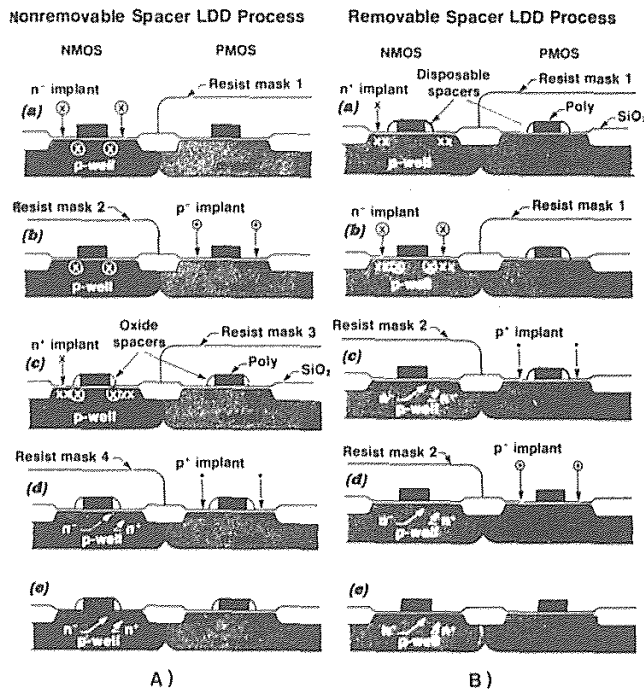


Fig. 6-38 (A) Non-removable oxide spacer LDD process: (a) n^+ implant, (b) p^+ implant, (c) after LDD spacer etch and n^+ implant, (d) after n^+ and p^+ implant, (e) after p^+ anneal. (B) Removable spacer LDD process: (a) after spacer etch and n^+ implant, (b) after NMOS spacer removal and n^+ implant, (c) p^+ implant, (d) after PMOS spacer removal and p^+ implant, (e) after p^+ anneal.⁶⁷ (© 1986 IEEE).

CVD oxide (over a thin polysilicon film) is used in the other.⁷³

Various techniques have been developed for forming the shallow source/drain junctions needed for submicron CMOS devices. First, As is implanted for the n -channel devices and BF_2^+ for the p -channel devices, since both species have very shallow projected ranges at implant energies of 30-50 keV (see Vol. 1, chap. 9). These implants are usually performed through a screen oxide to protect the source/drain regions from contamination during the implant procedure. Second, preamorphization of the silicon by implantation with Si or Ge reduces channeling and helps produce shallow junctions. It is necessary, however, to diffuse the implanted species past the layer of implant damage that cannot be annealed out, in order to prevent junction leakage (see Vol. 1, chap. 9). RTP techniques have been explored as a means of carrying out these anneal and diffusion thermal cycles. Shallow p^+n junction formation through the use of

diffusion (i.e., by utilizing RTP and either a solid diffusion source⁸² or a spin-on diffusion source⁸³) has also been reported. The use of an antimony implant to obtain shallow n^+p junctions with n^+ layers of lower resistivity than is possible with arsenic has also been described.⁸¹

Note that the screen oxide is usually stripped off following the source/drain implant, since it is damaged (and may be contaminated) by the heavy implants. A new oxide is grown over the source/drain regions and on the sidewall of the etched polysilicon electrode (and this step is referred to as *poly reoxidation*). During this thermal cycle, some (but not all) of the implantation damage is annealed out. The oxide formed on the poly sidewall shifts the remaining damage away from the gate edge to maintain the integrity of the thin gate oxide (Fig. 6-39).^{104,110} At the same time, the sidewall oxidation produces a *gate bird's beak*, or GBB, under the polysilicon edge. This reduces the gate-to-drain overlap capacitance and relieves the electric-field intensity at the corner of the gate structure. However, because the GBB encroachment can degrade transconductance of submicron MOSFETs and impact their subthreshold swing and threshold voltage,¹⁰⁹ this poly reoxidation process must be optimized for fabrication of submicron MOSFETs.¹¹¹

An approach to forming shallow junctions that uses CoSi_2 source/drain junctions has been reported.²⁹ In this approach, the CoSi_2 is formed before the source and drain junctions are created by means of a heavy ion implantation step. The implant is then performed so that the damaged layer, which would have occurred in the silicon crystal, is kept within the CoSi_2 layer. As a result, it is necessary to drive the implanted impurities just far enough so that they enter the silicon region to form the required pn junction. If a salicide process is used, the device has a cross-section, as shown in Fig. 6-40.

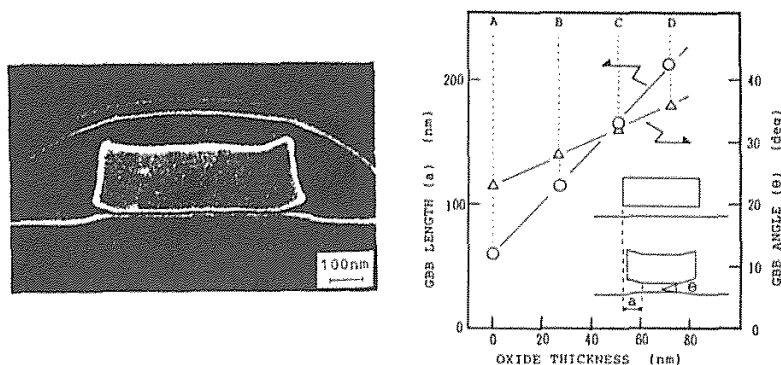


Fig. 6-39 (a) SEM micrograph of a gate structure with a GBB. (b) Profiles of GBB versus oxide thickness grown during re-oxidation.¹⁰⁴ This paper was originally presented at the Spring 1989 Meeting of The Electrochemical Society, Inc. held in Los Angeles, CA.

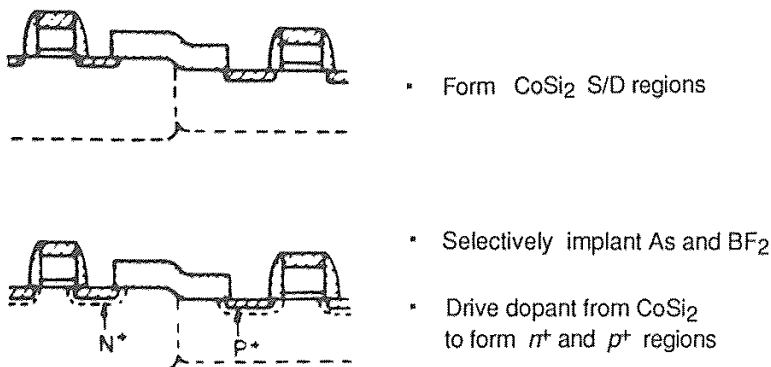


Fig. 6-40 Cross section of CMOS structure in which CoSi₂ is first formed on the source and drain regions, and then As and BF₂⁺ is selectively implanted into the CoSi₂. These dopants are then driven out of the CoSi₂ and into the Si substrate to form shallow p⁺ and n⁺ source/drain regions.²⁹ (© 1986 IEEE).

Another approach is to create so-called *elevated source-drains*. A thin (e.g., 200-nm) epitaxial layer of silicon can be selectively deposited onto the exposed source-drain areas of the MOS transistor, following the implantation of the lightly doped region of the LDD structure and formation of the spacers (Fig. 6-41a). This method has been used for the NMOS devices of a 1 Mbit SRAM.⁷⁹ In this case, a heavy BF₂⁺ implant is done such that the SEG film becomes heavily doped; the boron penetrates to the substrate during an RTP anneal that follows the implant. In this way, elevated, heavily doped, shallow source/drain regions are formed. The source-drain junction depths in the substrate are less than 0.2 μm deep (Fig. 6-41b). As noted earlier, the gate oxide that covers the source and drain regions is usually etched away and regrown following the implant step.

6.6.2.7 CVD Oxide Deposition and Contact Formation. Following the formation of the source and drain regions (and the salicide layers), a doped oxide is deposited by CVD, with procedures very much like those of the basic NMOS and CMOS processes (see section 5.4.1.6).

Contact windows are opened in this CVD layer to allow electrical connections to be made between the Metal 1 layer and the source/drain, gate, and substrate and well contact regions (*Mask #6*). Again, the details are similar to those of the basic NMOS and CMOS processes.

Advances in contact technology for CMOS include the use of barrier layers to prevent spiking through the shallow junctions, and the filling of contact holes by such materials as CVD W, polysilicon, or even SEG. In addition, as device dimensions decrease, the parasitic resistances of the source and drain regions become more significant. More details on all of these topics are provided in chapter 3.

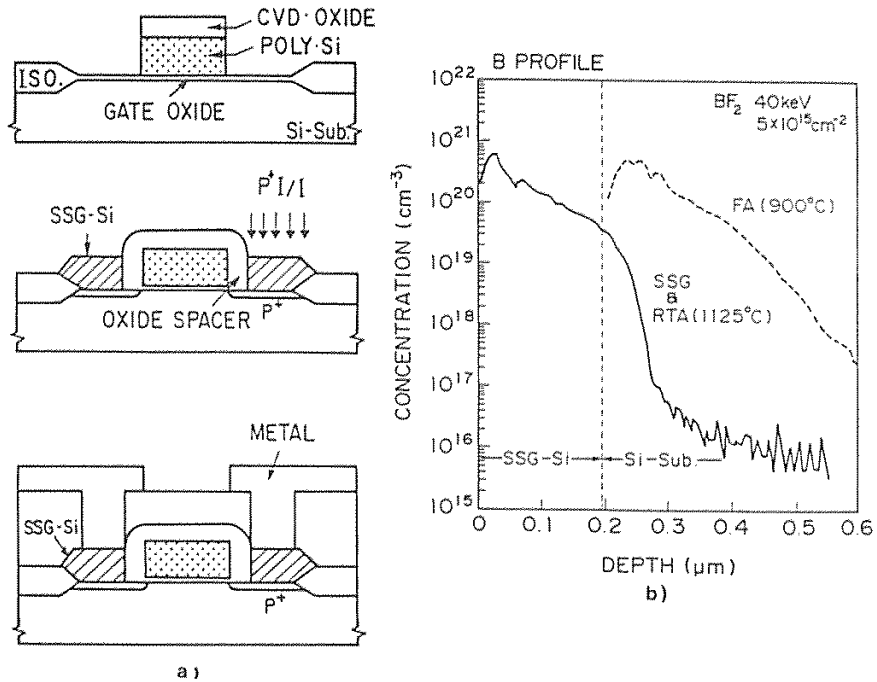


Fig. 6-41 Elevated source/drains. (a) Key process steps in forming elevated source/drains using SEG. (b) Comparison of the doping profile in source/drain regions of a PMOS device after a BF_2^+ implant directly into the regions (and a conventional furnace anneal) versus a BF_2^+ implant into the SEG regions and an RTP anneal.⁷⁹

6.6.2.8 Metal 1 Deposition and Patterning. The issues involved in the deposition and patterning of the Metal 1 layer are not significantly different from those described in chapter 5 for the NMOS process. *Mask #7* is used to pattern the Metal 1 layer. (For more information on the metal-layer deposition and patterning processes, see chapter 4 of this volume, and Vol. 1, chap. 9.)

Despite its higher resistivity compared to Al, tungsten's superior electromigration properties make it more appropriate for certain applications. As a result, CVD W has been selected as the Metal 1 material for a variety of circuit designs.

6.6.2.9 Intermetal Dielectric Deposition/Planarization and Via Patterning. Following the Metal 1 patterning, an intermetal dielectric must be deposited to electrically isolate Metal 1 from the Metal 2 layer. A variety of materials and deposition processes, have been utilized (see chap. 4).

However, deposition of this layer may make the wafer topography too severe to allow the Metal 2 film to be deposited with adequate step coverage. A number of *planarization* processes have been developed to overcome this problem (see chap. 4). If such techniques are successfully implemented, the wafer topography will be relatively planar, and any steps on its surface will be gently sloped rather than severely vertical.

It is necessary to open vias in the intermetal dielectric layer so that an electrical connection can be established between Metal 2 and Metal 1 at desired locations (*Mask #8*). Metal 2 must be deposited with adequate step coverage into the vias.

A number of techniques have been studied to improve coverage of Metal 2 in the vias, including the following (see chap. 4):

- Sloping the via walls by means of the via-etch process.
- Filling the vias with a blanket W or polysilicon deposition, and then etching back to provide a planar surface.
- Filling the vias with a selective deposition of W or Al.
- Increasing the step coverage into the vias through bias sputtering and heating of the substrate during deposition.
- Laser melting the Metal 2 film to increase step coverage.

6.6.2.10 Metal 2 Deposition and Patterning. The processing issues of depositing and patterning (*Mask #9*) the Metal 2 layer are discussed in chapter 4.

6.6.2.11 Passivation Layer Deposition and Patterning. The passivation-layer deposition and patterning (*Mask #10*) issues are the same those of the basic CMOS or NMOS processes.

6.7 MISCELLANEOUS CMOS TOPICS

6.7.1 Electrostatic-Discharge Protection

The input signals to an MOS IC are fed to the gates of MOS transistors. If the voltage applied to the gate insulator becomes excessive, the gate oxide can break down. The dielectric breakdown strength of SiO_2 is approximately 8×10^6 V/cm; thus, a 15-nm gate oxide will not tolerate voltages greater than 12 V without breaking down. Although this is well in excess of the normal operating voltages of 5-V integrated circuits, voltages higher than this may be impressed upon the inputs to the circuits during either human-operator or mechanical handling operations.

The main source of such voltages is triboelectricity (electricity caused when two materials are rubbed together). A person can develop very high static voltage (i.e., a few hundred to a few thousand volts) simply by walking across a room or by removing an integrated circuit from its plastic package, even when careful handling procedures are

followed. If such a high voltage is accidentally applied to the pins of an IC package, its discharge (referred to as *electrostatic discharge*, or ESD) can cause breakdown of the gate oxide of the devices to which it is applied. The breakdown event may cause sufficient damage to produce immediate destruction of the device, or it may weaken the oxide enough that it will fail early in the operating life of the device (and thereby cause device failure). A more detailed description of ESD failures in semiconductor devices can be found in reference 84.

All pins of MOS ICs must be provided with protective circuits to prevent such voltages from damaging the MOS gates. The need for such circuits is also mandated by the increasing use of VLSI devices in such high-noise environments as personal computers, automobiles, and manufacturing control systems. These protective circuits, normally placed between the input and output pads on a chip and the transistor gates to which the pads are connected, are designed to begin conducting or to undergo breakdown, thereby providing an electrical path to ground (or to the power-supply rail). Since the breakdown mechanism is designed to be nondestructive, the circuits provide a normally open path that closes only when a high voltage appears at the input or output terminals, harmlessly discharging the node to which it is connected.

Four types of circuits are used to provide protection against ESD damage, as follows:

1. diode breakdown
2. node-to-node punchthrough
3. gate-field-induced breakdown
4. parasitic *pnpn* diode latchup.

Often, a combination of protection methods is used, with a breakdown diode and one of the other protection devices connected in parallel with the gate being protected.

6.7.1.1 Diode Protection. Protection is obtained by using the diode-breakdown phenomenon to provide an electrical path in the silicon substrate that consists of a diffused resistor region (of a doping type opposite to that of the substrate). This diffused region is connected between the input pad and the gate (Fig. 6-42a). If a reverse-bias voltage greater than the breakdown voltage of the resultant *pn* junction is applied, the diffusion region (which otherwise works as a resistor), operates as a diode and undergoes breakdown. Furthermore, this diffused region will also clamp a negative-going transition at the chip input to one diode drop below the substrate voltage. In CMOS technologies, an additional protection diode can be added by utilizing the *pn* junction that exists between a p^+ node and the body region of the PMOS device (an *n*-type region, that is connected to V_{DD}). This diode is utilized as a protection device when a connection is made between the pad and a p^+ region. (Note that this second diode will clamp positive-going transitions to one diode drop *above* V_{DD} .)

6.7.1.2 Node-to-Node Punchthrough. As defined in section 5.5.2, *punchthrough* is the phenomenon by which the depletion region surrounding the drain of an MOS device extends along the channel of the device and contacts the depletion region of the source of the device. While the current that results from punchthrough is ordinar-

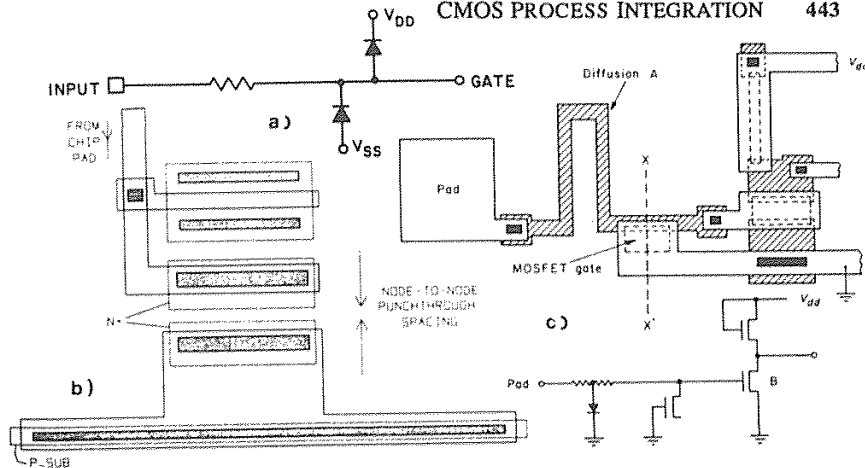


Fig. 6-42 MOS protective circuits: (a) Diode; (b) Punchthrough device; (c) Layout of an input protection structure containing both devices.¹²² From T. E. Dillinger, *VLSI Engineering*. Copyright Prentice-Hall, 1988. Reprinted with permission.

arily an unwanted effect, it is harnessed in ESD protection circuits for beneficial purposes. The source and drain regions of an ESD punchthrough protection structure are placed close enough together that punchthrough occurs at a voltage lower than that of the gate-oxide breakdown (Fig. 6-42b). Such a structure is laid out and tested to meet a specific ESD transient reliability measure. Once it has been verified as functioning according to the design specifications, it can be connected to each chip signal I/O pad.

6.7.1.3 Gate-Controlled Breakdown Structure. The electric field near the corner of an MOS device's drain node is strongest at the surface, since the depletion region is narrowest at this point⁸⁵ and the entire voltage across the depletion region must be dropped over a very short distance. The fact that the gate voltage can increase the strength of the electric field at the corner of the drain is the principle used in designing the gate-controlled breakdown structure (Fig. 6-43).

As the reverse-bias voltage applied to the drain is increased, the gate-to-drain voltage also increases (note that the gate is tied to ground). Because the presence of the gate reduces the breakdown voltage of the junction near the corner of the drain, a relatively small voltage can cause the junction to break down at this point. At breakdown, the junction conducts a large current; in this device, the current flows from the drain to the substrate. The actual input voltage at which breakdown occurs can be controlled by altering the oxide thickness. This produces a large voltage drop across the relatively high resistance of the diffusion area, reducing the voltage applied to the gate of the MOSFET.

Such structures were the workhorses of ESD protection devices. It has been found, however, that processing enhancements that increase the device performance of small-dimension devices sometimes have a negative impact on the failure resistance of the

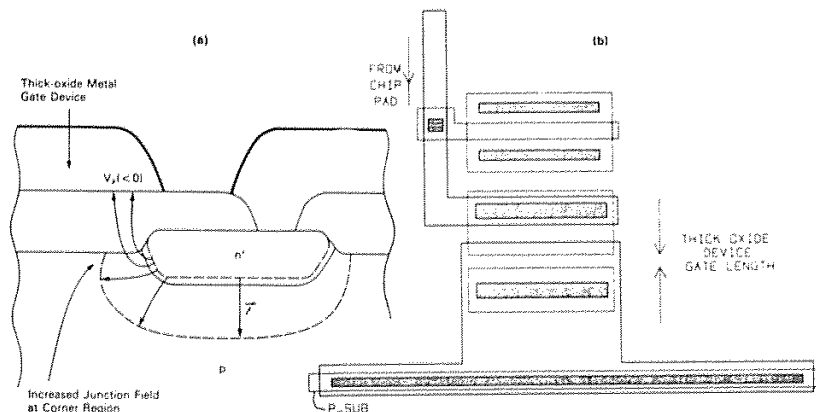


Fig. 6-43 (a) Cross-section, and (b) layout of a thick-oxide metal-gate device for node-to-substrate breakdown and input-pad protection.¹²² From T. E. Dillinger, *VLSI Engineering*. Copyright Prentice-Hall, 1988. Reprinted with permission.

ESD protective devices. Specifically, silicided source/drain regions and LDD structures can degrade the ESD performance of the gate-controlled breakdown structures so that the voltage at which *these devices themselves fail* is reduced from 4 kV to 1 kV.^{86,87} As a result, ESD protection structures can be rendered largely ineffective by the silicided regions, or else the circuits can be caused to fail by ESD in new failure modes as a result of the addition of performance-enhancement features.^{86,88}

6.7.1.4 *pnpn*-Diode ESD Protection for Advanced CMOS Circuits.

Even as devices are made smaller, a fixed minimum volume of silicon must always be used to dissipate the power associated with the current flowing through the protective devices to ground. If the power is dissipated in too small a volume of silicon, the silicon can be heated beyond its melting point; this can destroy the device, even if the circuit-protection device is operating properly. As a result, a minimum unscalable area (Fig. 6-44) will be required in order for the same degree of protection to be maintained through each generation of scaled technology.^{68,89}

Effective protection of VLSI must be provided for both automated and human handling. ESD from a machine is typically modeled as a high-voltage (4-kV), short-duration pulse (i.e., the *EIAJ machine model [MM] test method*), while ESD from a human operator is usually modeled as a longer, lower-voltage (300-V) but higher-current pulse (e.g., the *Mil. Std. human body model [HBM] test method*). Thus, if the gate-controlled breakdown structure itself fails at 1 kV, it ceases to provide protection against HBM ESD.⁸⁷

To solve this problem on the inputs of a CMOS chip, it has been proposed that the parasitic lateral *pnpn* structure inherent in all CMOS circuits be exploited (Fig. 6-45a)

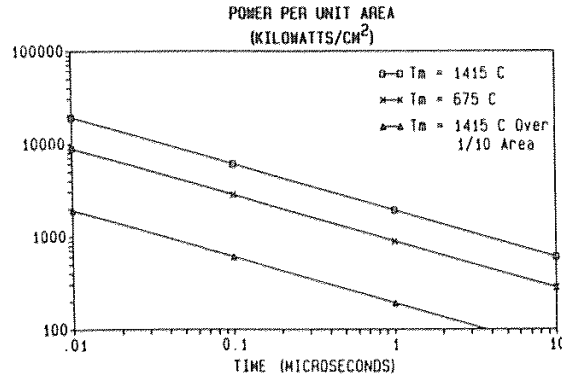


Fig. 6-44 Wunsch and Bell power curves for silicon melt, aluminum melt, and silicon melt for nonuniform current density.⁸⁹ (© 1968 IEEE).

as an ESD structure (see section 6.4). Such a structure – which is deliberately configured to latch up at a voltage lower than that required to damage the input MOS gate oxide – is built into each input circuit on the chip (D1 in Fig. 6-45b). Thus, this parasitic *pnpn* structure serves as the primary ESD structure of the input circuit by

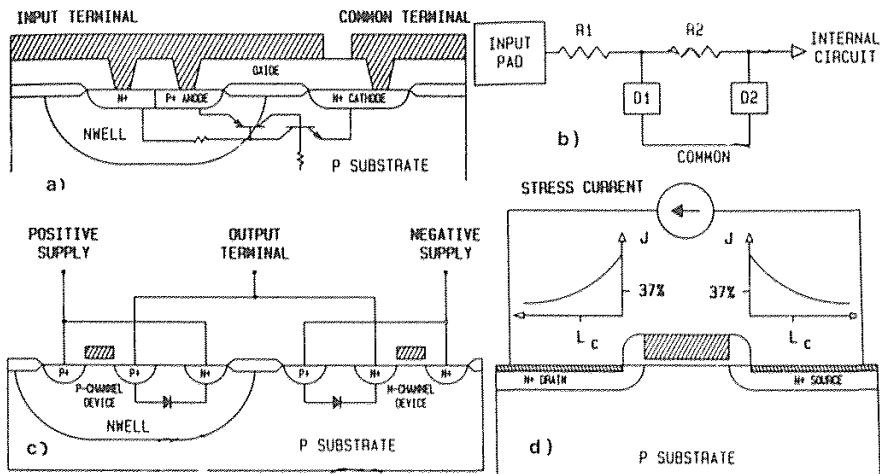


Fig. 6-45 (a) Lateral *pnpn* input protection structure showing parasitic and increased area at *n*-well and *p*-well junction resulting in lower power density. (b) Input-protection circuit showing primary protection elements R1 and D1 and secondary protection elements R2 and D2. (c) Representation of dual-diode output circuit cross section. Effectiveness depends on parasitic diode impedance as well as capacitance between supplies. (d) Cross section of MOS thin-oxide device showing current density at silicide and silicon interface.⁸⁷ (© 1979 IEEE).

providing a low-impedance path when it is turned on. It also offers increased area for power dissipation. Secondary ESD structures, however, must also be used on each input circuit (e.g., D2 and R2 in Fig. 6-45b) to provide effective protection until D1 achieves its low-impedance conducting state.

On the outputs of the chip, the parasitic diodes that exist in the CMOS output circuit are used to provide ESD protection, much as described in section 6.7.1.1 (Fig. 6-45c).⁸⁷ Such dual-diode CMOS outputs have proved to be effective when the resistance of the parasitic diode in the *p*-channel device is minimized and the failure threshold of the *n*-channel device is maximized. As shown in Fig. 6-45d, the stress current density in the diode of the silicided *n*-channel device is very nonuniform due to the silicided source/drain regions. As a result, the diode can fail if too high a current density passes through only a small fraction of the area of the silicided contact.

6.7.2 Power-Supply Voltage Levels for Future CMOS

As MOS has evolved, device scaling has been practiced to increase the speed and packing density of MOS ICs. Although scaling of the electric field to maintain a constant electric field (CE) would have promoted both high performance and high reliability, in practice a constant voltage (CV) has been maintained in order to maintain TTL compatibility and produce higher circuit performance. Such nonscaling of the voltage leads to reliability problems, caused by hot-electron effects and oxide breakdown. Scaling of the power supply voltage is thus desirable. M. Kakamu has calculated the optimum power-supply voltage as a function of shrinking line width, and has shown that this voltage can be scaled down in proportion to the square root of the design-rule shrinkage without sacrificing circuit performance.⁹⁰ The relationships given between the supply voltage and the design rule are

$$V_{DD} = 6.1 \times (L/2)^{1/2} \text{ (for conventional drain structures)} \quad (6-9)$$

and

$$V_{DD} = 8.2 \times (L/2)^{1/2} \text{ (for LDD structures)} \quad (6-10)$$

where L (μm) is the design rule and V_{DD} is the power-supply voltage (volts). These equations predict that at dimensions of $0.6 \mu\text{m}$, the power supply voltage will have to be reduced to 3.3 V in order for reliable device operation and circuit performance to be maintained if conventional drain structures are used. On the other hand, if LDD structures are used, it is predicted that $V_{DD} = 5 \text{ V}$ will still be feasible for $0.7\text{-}\mu\text{m}$ dimensions. A recent report indicates that these equations may be slightly pessimistic, in that MOSFETs with gate lengths of $0.6 \mu\text{m}$ that operate from a 5-V supply have been successfully fabricated.¹⁰² Through optimization of the LDD structures of these devices, a lifetime of more than 10 years in the face of hot-carrier degradation effects, is predicted.

6.7.3 Low-Temperature CMOS

Since it is possible to build MOS devices with minimum dimensions of less than 0.5 μm , low-temperature operation of CMOS (e.g., at 77°K) has been actively investigated. At low temperatures, the MOS devices exhibit lower subthreshold leakage, higher carrier mobility (which yields improved speed performance), and a steeper logarithmic current-voltage slope. It has been found that the normalized propagation delay in CMOS logic gates is reduced by a factor of nearly 2 or 3 when the devices are operated at 77°K.^{91,92,93,94} The major reason for the slow commercial introduction of such systems has been the difficulty associated with liquid-nitrogen refrigeration.

6.7.4 Three-Dimensional CMOS

Since the limits of two-dimensional CMOS may soon be approached, a logical direction for continued advances in device scaling is three-dimensional devices. CMOS is well-suited for 3-D integration, since low-power circuits will be mandatory in order for heat dissipation problems to be avoided. The 3-D approach will offer the benefits of increased packing density and higher speeds (due to shorter interconnects).

The inability to fabricate single-crystal silicon over insulating layers, however, has historically been an obstacle to the implementation of 3-D ICs. Recent advances in SOI technologies show promise of being able to surmount this obstacle and much work is currently being done to develop 3-D CMOS.⁹⁵

REFERENCES

1. J. Y. Chen, *CMOS Devices and Technology for VLSI*, Englewood Cliffs N.J., Prentice-Hall, 1989.
2. F. M. Wanlass and C. T. Sah, *IEEE Int. Solid-State Circ. Conf.*, February 1963.
3. R. D. Davies, "The Case for CMOS", *IEEE Spectrum*, October 1983, p. 26.
4. J. Y. Chen, "CMOS-The Emerging Technology," *IEEE Circuits and Systems Magazine*, March 1986, p. 16.
5. D. M. Brown, M. Ghezzi, and J. M. Pimbley, "Trends in Advanced CMOS Process Technology," *Proceedings of the IEEE*, December 1986, p. 1678.
6. W. C. Holton and R. K. Calvin, "A Perspective on CMOS Technology Trends," *Proceedings of the IEEE*, December 1986, p. 1646.
7. S. Prussin, private communication.
8. R. Chwang and K. Yu, "CHMOS-An n -Well Bulk CMOS Technology for VLSI," *VLSI Design*, Fourth Quarter, 1981, p. 42.
9. W. Wijaranakula et al., *J. Electrochem. Soc.*, December, 1988, p. 3113.
10. H. Tsuya et al., *Japanese J. Appl. Phys.*, 22, L16, (1983).
11. T. Yamaguchi et al., *IEEE Trans. Electron Dev.*, ED-31, 205 (1984).
12. L. C. Parrillo, "CMOS Active and Field Device Fabrication," *Semicond. Internat.*, April 1988, p. 64.

13. A. G. Lewis et al., "Vertical Isolation in Shallow n -Well CMOS Circuits," *IEEE Electron Dev. Letts.*, March 1987, p. 197.
14. L. C. Parrillo et al., "Twin-Tub CMOS," *IEDM Tech. Dig.*, 1980, p. 752.
15. J. Manoliu and J. O. Borland, "A Submicron Buried-Layer Twin-Well CMOS SEG Process," *IEDM Tech. Dig.*, 1987, p. 20.
16. R. D. Rung, C. J. Dell' Oca, and L. G. Walker, "A Retrograde p -Well for High-Density CMOS," *IEEE Trans. Electron Dev.*, **ED-28**, p. 1115, 1981.
17. R. A. Martin and J. Y. Chen, "Optimized Retrograde n -Well for 1 μm CMOS," *Proc. Custom Integ. Circ. Conf.*, 1985, p. 199.
18. S. R. Combs "Scalable Retrograde p -Well CMOS Technology," *IEEE Trans. Electron Dev.*, **ED-28**, p. 346, 1981.
19. Y. Taur et al., *J. Solid-State Circuits*, **SC-20**, p. 123, 1985.
20. S. M. Sze, Ed., *VLSI Technology*, 2nd. Ed., New York, McGraw-Hill, 1988, p. 491.
21. T. Ohzone et al., *IEEE Trans. Electron Dev.*, **ED-32**, 1789, (1980).
22. K. M. Cham, S.-Y. Oh, D. Chin, and J. L. Moll, *Computer-Aided Design and VLSI Device Development*, Boston, Kluwer Academic Publishers, 1986 p. 182.
23. G. J. Hu et al., *IEDM Tech. Dig.*, 1982, p. 710.
24. J. Zhu et al., *IEEE Trans. Electron Dev.*, February 1988, p. 145.
25. A. Schmitz and J. Y. Chen, "Design, Modelling, and Fabrication of Submicron CMOS Transistors," *IEEE Trans. Electron Dev.*, **ED-33**, p. 148, (1986).
26. R. A. Chapman et al., "A 0.8 μm CMOS Technology," *Tech. Dig. IEDM*, 1987, p. 362.
27. L. C. Parrillo et al., *Tech Dig. IEDM*, 1984, p. 418.
28. K. M. Cham et al., *IEEE Electron Dev. Lett.*, January 1986, p. 49.
29. S. J. Hillenius et al., "A Symmetric Submicron CMOS Technology," *Tech. Dig. IEDM*, 1986, p. 252.
30. Y. Pauleau, "Interconnect Materials for VLSI Circuits, Part-1," *Solid State Technology*, February 1987, p. 61.
31. V. Schwabe, F. Neppl, and E. P. Jacobs, *IEEE Trans. Electron Dev.*, **ED-31**, 1984, p. 988.
32. S. Iwata et al., *IEEE Trans Electron Dev.*, **ED-31**, 1984, p. 1174.
33. R. F. Kwasnick et al., *IEEE Trans. Electron. Dev.*, September 1988, p. 1432.
34. H. Oikawa and T. Amazawa, *Proc. 3rd. Intl. VLSI Symp., ECS Meeting*, May, 1985, p. 131.
35. D. M. Brown et al., *IEEE Trans. Electron Dev.*, **ED-18**, 1971, p. 931.
36. R. R. Troutman, *Latchup in CMOS Technology*, Kluwer, Boston, MA., 1986.
37. G. J. Hu, "A Better Understanding of CMOS Latchup," *IEEE Trans. Electron Dev.*, **ED-31**, p. 62, 1984.
38. R. R. Troutman, "Latchup in CMOS Technologies," *IEEE Circuits & Sys. Mag.*, May 1987, p. 15.
39. A. G. Lewis et al., *Tech. Dig. IEDM*, 1986, p. 248.
40. R. D. Rung and H. Momose, *IEEE Trans. Electron Dev.*, **ED-30**, 1983, p. 1647.
41. W. R. Dawes and G.F. Derbenwick, "Prevention of CMOS Latchup by Gold-Doping," *IEEE Trans. Nucl. Sci.*, **NS-23**, 1976, p. 2027.

42. J. R. Adams and R. J. Sokel, Presented at the 1979 Nuclear and Space Radiation Effects Conf., Santa Cruz, CA., July, 19, 1979.
43. J. O. Borland and T. Deacon, "Advanced CMOS Epitaxial Processing for Latchup Hardening," *Solid-State Technology*, p. 123, August 1984.
44. S. Ratanphanyarat et al., *Tech. Dig. IEDM*, 1987, p. 744.
45. F. S. Lai et al., *IEDM Tech. Dig. 1985*, p. 513.
46. M. L. Chen et al., *Tech. Dig. IEDM*, 1986, p. 256.
47. S. Swirhun et al., "Latchup-Free CMOS Using Guarded Schottky Barrier PMOS," *Tech. Dig. IEDM*, 1984, p. 402.
48. R. A. Martin et al., *IEDM Tech. Dig.*, 1985, p. 403.
49. A. G. Lewis, *IEEE Trans. Electron Dev.*, Oct. 1984, p. 1472.
50. D. Takacs et al., *Tech. Dig. IEDM*, 1983, p. 159.
51. K. W. Terrill et al., *Tech. Dig. IEDM*, 1984, p. 406.
52. R. Menozzi et al., "Layout Dependence of CMOS Latchup," *IEEE Trans. Electron Dev.*, Nov. 1988, p. 1892.
53. J. Y. Chen and D. E. Snyder, "Modeling Device Isolation in High-Density CMOS," *IEEE Electron Dev. Letts.*, February 1986, p. 64.
54. L. Herman, "Controlling CMOS Latchup," *VLSI Design*, April, 1985, p. 100.
55. M. -L. Chen et al., *Tech. Dig. IEDM*, 1986, p. 256.
56. K. M. Cham and S.-Y. Chiang, "A Study of Trench Surface Inversion Problem for the Trench CMOS Technology," *IEEE Electron Dev. Lett.*, Sept., 1983, p. 303.
57. Y. Nitsu et al., "Latchup Free CMOS Structure Using Shallow Trench Isolation," *Tech. Dig. IEDM*, 1985, p. 509.
58. R. D. Rung, H. Momose, and Y. Nagakubo, "Deep Trench Isolated CMOS Devices," *IEDM Tech. Dig.*, 1982, p. 6.
59. T. I. Kamins and S. Shiang, *Electron Dev. Letts.*, December 1985, p. 617.
60. N. Kasai et al., *IEEE Trans. Electron Dev.*, June 1985, p. 1331.
61. A. Stivers et al., *10th Internat. Conf. on CVD, ECS*, 87-8, p. 389.
62. R. D. Rung et al., *IEEE Trans. Electron Dev.*, ED-28, 1981, p. 1115.
63. L. C. Parrillo et al., "Twin-Tub CMOS II," *IEDM Tech. Dig.*, 1982, p. 706.
64. S. J. Hillenius and L. C. Parrillo, U.S. Patent No. 4,554,726, Nov. 25, 1985.
65. J. Pfiester and J. Alvis, "Improved CMOS Field Isolation Using Ge/B Implantation," *IEEE Elect. Dev. Lett.*, August 1988, p. 391.
66. J. Pfiester, *Electron Dev. Letts.*, November 1988, p. 561.
67. L. C. Parrillo et al., *Tech. Dig. IEDM*, 1986, p. 244.
68. M. H. Woods, "MOS VLSI Reliability and Yield Trends," *Proceedings of IEEE*, Dec. 1986, p. 1715.
69. G. S. Ohrlein, *J. Appl. Physics*, 59, 1587 (1986).
70. B. W. Shen et al., *Tech. Dig. IEDM*, 1987, p. 582.
71. M. Saitoh, T. Mori, and H. Tamura, *Tech. Dig. IEDM*, 1986, p. 680.
72. K. Ogiue et al., "Technology Improvement for High Speed ECL RAMs," *Tech. Dig. IEDM*, 1986, p. 468.
73. J. Pfiester, *IEEE Electron Dev. Letters*, April, 1988, p. 189.
74. R. de Werdt et al., *Tech. Dig. IEDM*, 1987, p. 532.