Chapter 7

ADVANCED LITHOGRAPHY II: OPTICAL HARDWARE

In this chapter the topics related to microlithography optics and hardware pertinent to IC technology at the start of 2001 decade are discussed. These subjects include: excimer-laser light-sources; optical-alignment-tools (stepper-scanners) for the 248-, 193-, and 157-nm lithography generations; resolution-enhancement-techniques; and next-generation lithography (NGL).

7.1 EXCIMER-LASER DEEP-UV LIGHT SOURCES

Excimer lasers are used to produce the light needed for deep-UV optical lithographic exposure systems. Deep-UV laser sources will dominate what many industry experts believe will be optical lithography's final arena before an industry jump to the post-optical era. Such laser-light sources are essential enabling elements for lithography for the next three critical-dimension nodes of the industry's roadmap. The term *laser* is an acronym for *light amplification by stimulated emission of radiation*. The krypton-fluoride (KrF) excimer laser is used to produce DUV light with a wavelength of 0.248-μm (or 248-nm). Argon-fluoride (ArF) excimer lasers are employed to produce DUV light with a wavelength of 193-nm, and F_2 excimer lasers for 157-nm DUV light. No fundamental changes are necessary to allow a KrF-laser to operate as an ArF-laser, except to change the gas composition in the laser's plasma tube and the mirrors that make up the cavity.

By 2001 the KrF-excimer-laser had become the light source of choice in mainstream, high-volume, advanced IC production (i.e., for 0.35-, 0.25-, and 0.18-μm-CMOS technologies). At the same time, ArF-excimer-lasers were being installed as the light source for 193-nm stepper/scanners that were being used for advanced process development applications, as well as for some pilot-line production (i.e., for sub-0.2-μm-CMOS process-development lines). It is conjectured that ArF lasers will displace KrF between 130-nm and 100-nm, albeit not until ArF maturity is reached in production. ArF has the potential to be the dominant source down to the 100-nm to 70-nm technology-nodes. Below 100-nm, the F_2 laser appears as a possibility.[1]

An *excimer* is an exotic molecule formed by one atom each of a noble gas and halogen atom. This *dimeric* molecule is bound only in a quasi-stable excited state.

TSMC-1013
U.S. Patent 8,884,373

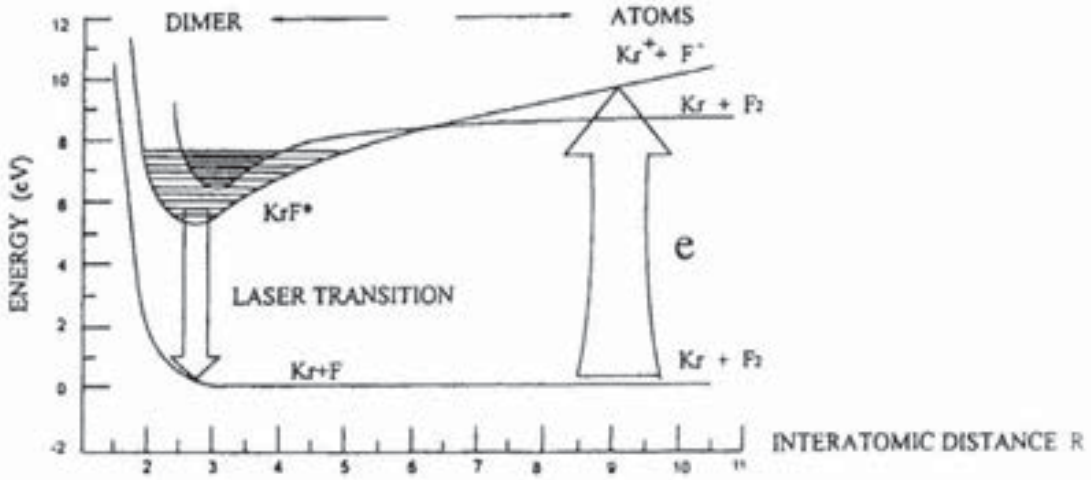


Fig. 7-1 Energy diagram for a KrF excimer laser. KrF^* is formed via two reaction channels. It decays to the ground state via dissociation into Kr and F while emitting a photon at 248-nm.

That is, if the noble gas and halogen gas atoms are in the ground state they will not react, but if one (or both) are in an excited state, a chemical reaction can occur to form a *dimer*. The term "excimer" is thus a contraction of the phrase "*excited dimer*." (Note that true dimer*s* are molecules containing two atoms of the *same* element, but this term is also used here to describe a molecule with two different atoms.) When the excimer decays back to the ground state, the dimer falls apart into its two constituent atoms (with the emission of a DUV-photon accompanying the decay process). When such decay is stimulated by absorption of energy from another light photon, DUV-laser light is emitted by an ensemble of KrF^* dimers. The spontaneous (i.e., not stimulated) dissociation time of an excimer (e.g., KrF^*) is relatively long (i.e., in the nanosecond to microsecond range). Therefore, population inversion (higher population of the excited state than the ground state) can readily be achieved by exciting a high-pressure mixture of Kr and F gases in a high-voltage pulse discharge (see Fig. 7-1).

A capacitor (that will serve as the transient power supply to produce the plasma discharge) is first charged to 12-kV. Then it is discharged in a very short time (75 ns) across electrodes present in the laser cavity. The plasma discharge created by this injection of energy produces Kr^+ and F^- ions, and these ions then combine to form the KrF^* excimers (with the process that creates the excited population of dimers called *pumping*). The laser light is produced by stimulating this created population of excimers to decay within a very short time. Some of the excimers begin to decay spontaneously. The light they emit passes through the gas containing the excimer population, and this stimulates additional excimer decay (which creates more radiation). This is the *amplification* process alluded to

in the acronym term *laser*. The laser cavity also uses mirrors to reflect radiation back into the gas. Thus, the stimulated decay rate (and laser-light emission) continues to increase, until almost all of the excimers have undergone stimulated decay. In fact, in a properly designed laser cavity, the stimulated decay process is completed very quickly after it is initiated. Thus, the light of an excimer laser is emitted in the form of short pulses of energy (i.e., these excimer-laser sources do not emit laser-light continuously, as do continuous-emission lasers). The laser-light exits the laser cavity though one of its ends, at which a partially-reflecting surface is located. As soon as the capacitor can be charged again to a high-enough voltage (making it able to act as the excitation source for creating the discharge) the excitation, pumping, and laser-emission cycle can be repeated.

One of the biggest problems in using excimer lasers as lithographic light-sources, however, is the fact they emit their energy in such a pulsed mode. Excimer lasers with a power output that is useful for lithographic applications (2-20 W) typically produce pulses of laser energy at a rate of 200-2500 Hz, with each pulse being only 5-20-nsec in length and containing about 10-mJ of energy. Since each pulse is so short, and there is a relatively long time between pulses, the peak power within each pulse is extremely high (even when the average power of the laser appears to be quite modest). Such peak powers can cause damage to the optical coatings and materials of the optical lithography system if it is concentrated in too small an area (e.g., the stepper lens). Thus, design features must be employed to keep the power-density low enough at all points within optical elements. An increase in the repetition rate of the laser is also beneficial, since this produces the same average power with a lower power per pulse (and exposure-dose-control is also improved).

Excimer lasers are attractive as DUV light-sources for several reasons. First, KrF-lasers can deliver much more usable power at 248-nm than can a mercury-arc lamp. The radiance of the laser is several orders-of-magnitude greater than the arc lamp because the laser emits a highly-collimated beam of light, whereas the arc lamp emits light in all directions (isotropically). Excimer-lasers also have a low degree of coherence compared to most other kinds of lasers, as well as a relatively broad linewidth (which results in low temporal coherence). Low coherence is generally a desirable characteristic of a lithographic light source.

On the other hand, the 0.300-nm (300-pm) *spectral width* of the 248-nm laser is not narrow enough to satisfy the bandwidth requirements of the refractive optical systems of step-and-repeat aligners. (That is, the dispersion [the change of optical index with wavelength] of optical components in the DUV region is high. Thus, spectral bandwidths even in the picometer range, can produce significant

chromatic aberration in systems with refractive optical components. This makes it difficult or even impossible to focus and image lithographic patterns properly.) Consequently, since the lenses of these steppers have no chromatic correction, this forces the illumination source to have a spectral bandwidth of less than 0.001-nm (or < 1-pm). It is therefore necessary to put dispersive optical elements (such as prisms and diffraction gratings) within the laser cavity to reduce the bandwidth of the KrF laser-light to this value. These elements somewhat reduce the total power of the laser, but more importantly they also decrease the stability of the power level. A feedback system must be employed to keep the center wavelength of the line-narrowed laser-light-beam constant to about the same level of accuracy as the bandwidth. This control is needed because if the laser wavelength drifts by a few thousandths of a nm, the chromatic aberrations of the lens will cause an excessive shift in the focus of the lens. For each shift of the wavelength by 1-pm (0.001-nm), the lens-focus will shift by about 0.15-μm.

7.1.1 KrF Excimer Lasers

By 2001, excimer-laser wavelength-stability of much less than 1-pm was being achieved (i.e., ±0.25-pm). For example, the Cymer ELS-6010 KrF laser (introduced in 2000) has a spectral bandwidth: ≤ 0.5-pm full-width-at-half maximum (FWHM – see Fig. 7-2a) and a 95% energy integral of ≤ 1.4-pm (this latter specification defines the spectral range that includes 95% of the energy of the pulse, see Fig. 7-2b). Also, the repetition rate of this laser has been increased to 2500 Hz. However, beyond 180-nm lithography, even a 0.5-pm spectral bandwidth and a 95% energy integral of 1.4-pm will not be sufficiently small. These specification will need to be narrowed to <0.4 pm and <1 pm, respectively. Using double-pass Etalon (reflecting surface, Fig. 7-4b) optics with a 10W, 2000

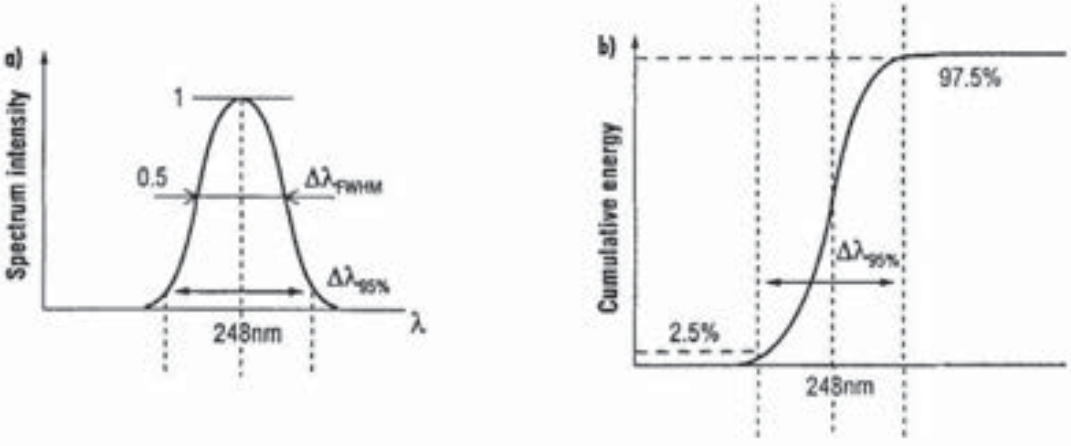


Fig. 7-2 a) Definition of "full width at half maximum," FWHM; b) Definition of 95% integral pulse bandwidth.

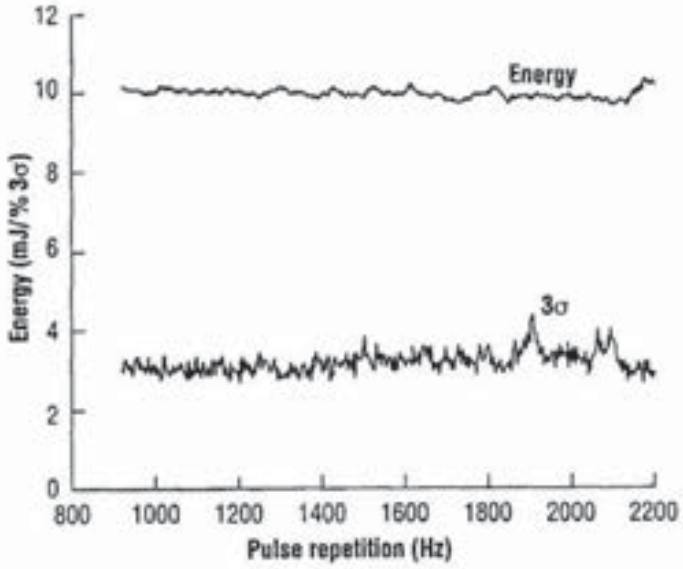


Fig. 7-3 Pulse energy and stability vs. repetition rate for a KrF laser running in the constant charge voltage mode.

Hz laser, laser pulses with an actual spectral bandwidth centered at 248.3-nm, a spectral bandwidth of 0.27-pm and 95% energy integral of 0.93 were produced.[2]

Another issue related to pulse power is exposure control. The method for controlling exposure in a stepper illuminated with a pulsed laser is to integrate the pulsed energy until the required exposure dose is reached (at which point a signal is sent to stop the laser from pulsing). To achieve a 1% control of accuracy, at least 100 pulses must be used. Early excimer lasers had lower pulse-repetition-rates (1000 Hz), and thus had slower exposure times. By doubling the pulse-repetition-rate to 2000 Hz, the delivered exposure energy within the same unit of time was also doubled, and this also cut the exposure time approximately in half. By 2002, the newest laser models have overcome this limitation (repetition-rates are up to 4000 Hz), and the exposure speeds of excimer laser steppers now match (or exceed) those of arc-lamp-illuminated systems. Photonic energy output is a related critical parameter for lithography. The recently improved output characteristics of KrF systems are such that the relationship between output power and laser repetition rate is essentially linear (Fig. 7-3).

In addition to the issue of pulsed power, excimer lasers have some other drawbacks. They are much larger than an arc-lamp illumination source, typically occupying a clean-room footprint of 2x5 feet (Fig. 7-4). The excimer laser is also expensive ($500,000–$1,000,000), which in 2001 added 10-20 percent to the cost of the stepper. Gas used by the stepper is also costly and toxic and the plasma cavity in which it is contained, is not sealed. Thus, it must be periodically refilled

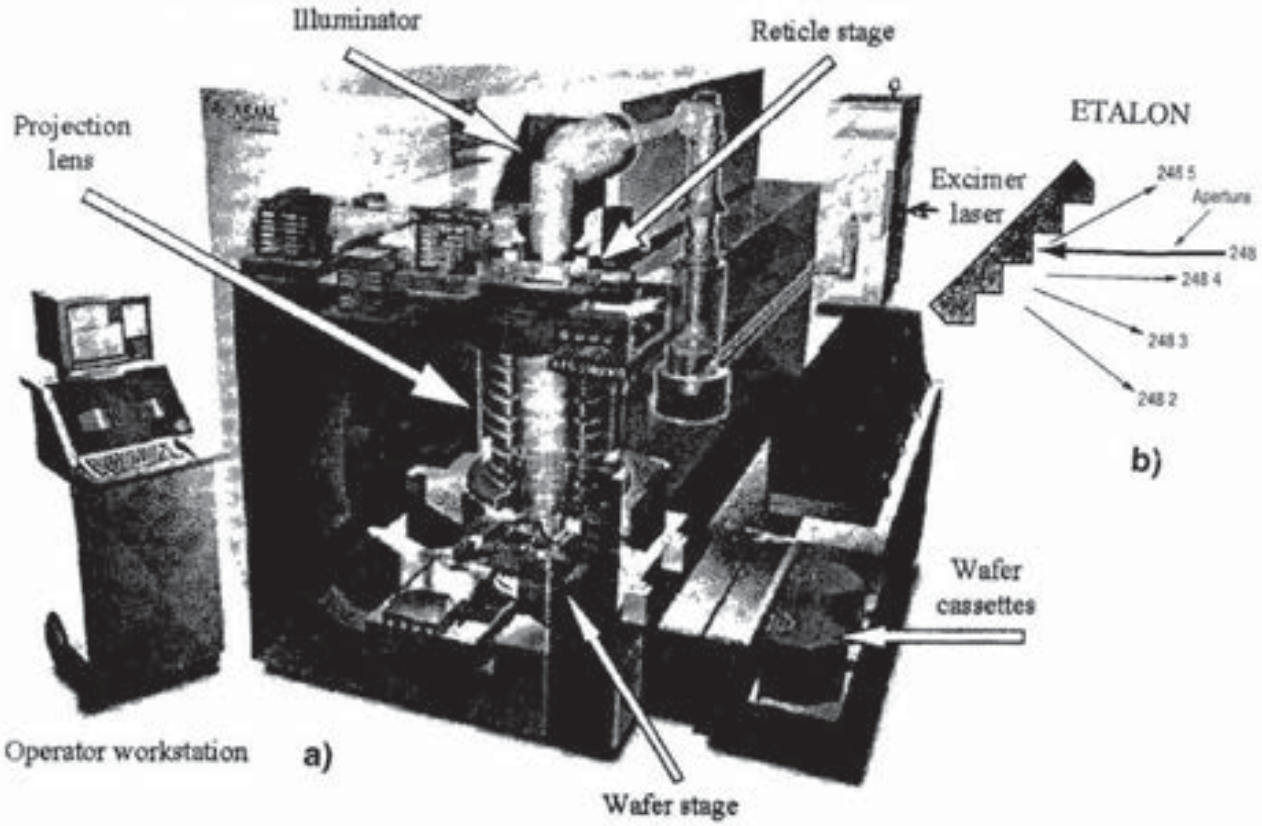


Fig. 7-4 a) Configuration of excimer laser light source and stepper (Courtesy ASML); b) Etalon

with this gas (about every 5 days). This requires a gas-handling system that can fulfill the rigorous industrial safety requirements for toxic gases. (The interval between gas refilling has been increased by more than a factor of 10 since the earliest lasers were used for lithography, and improvement is continuing.)

The laser pulses also sputter materials onto the mirrors and windows of the cavity. Thus, they must be cleaned after a number of pulses. The laser must be shut down during such servicing. Eventually, the entire laser-cavity must be replaced (which represents the largest cost of operating an excimer laser). The number of pulses between cleanings (and replacement) has been continuously increased. In 1998, servicing was required after 1 billion pulses and laser-cavity replacement after more than 2.5-billion pulses. Since laser usage during production is about $2x10^9$ pulses in a year, in 1998 window cleaning had be done about once every six months, and laser-cavity replacement about once a year.

The laser system is also relatively inefficient in terms of electrical-power usage, and thus it consumes about 10-kW while the laser is operating. Finally, the DUV-light of the laser is very hazardous to human skin and eyes. Hence, stringent safety precautions must be enforced when maintenance work is performed on the laser. The laser enclosure and beamline must also be interlocked to prevent anyone operating the system from being exposed to the

laser-light. Nevertheless, as excimer-laser technology has matured the reliability of lasers has been greatly increased, and the frequency-of-repair rates have dramatically declined. KrF lasers are commercially available from Cymer, Inc., Lambda Physik, and Komatsu, with Cymer, Inc. at the turn of the century having the largest market share.

7.1.2 ArF Excimer Lasers

The demands of ArF-laser technology are even more stringent than those of KrF-lasers. The power output and repetition-rate must be increased, line-narrowing must be smaller, and pulse-uniformity must be improved to allow accurate dose-control. Finally, the entire set of issues dealing with accelerated degradation of materials and system components caused by 193-nm radiation must be addressed. It appears that a basic space-line resolution of less than 105-nm will be possible using ArF illumination. By 2002, Cymer introduced the Nanolith-7000™ ArF-laser with a 4-kHz rep-rate, an average power of 20-W, and high spectral purity. The optical-component degradation under 193-nm radiation was also found to be lessened as the peak-power of the laser pulse is decreased. Thus, efforts to reduce the peak-power (by increasing the pulse-width) were being pursued for ArF-lasers. Some success toward achieving this goal have been described. Progress reported by 2000 indicates that the manufacturing performance of ArF-lasers will be comparable to that exhibited by KrF-lasers currently being used in production.

7.1.3 F_2 Excimer Lasers

There are many challenges that must be met to build production-worthy F_2 excimer lasers, but these also appear to be tractable. As with the ArF-laser, reducing peak power, but increasing average power (10-40-W) will be important. F_2-lasers operating at multi-kHz pulse rates will be needed. Spectral linewidth demands at 157-nm will depend on lens design, but it seems likely that a spectral bandwidth of 0.2-pm will be required. F_2-lasers emit two bands of light (at 157.63-nm and 157.52-nm), but 87% of the energy is located at the longer wavelength. A prototype 2000-Hz F_2-laser, based on a current ArF design was described in 1999, in which single-line operation at 157.63-nm was achieved. The output of this F_2-laser scaled linearly with repetition rate. Its energy stability was better than 6% (3σ) over 100 bursts and better than 4% over more than 80% of the burst range. Gas lifetimes appear comparable to current ArF lifetimes.

7.2 EXPOSURE TOOLS FOR DUV LITHOGRAPHY

The wafer stepper was initially commercialized by the GCA Corporation of Andover, MA. GCA's first stepper, the DSW4800 (Direct Step on the Wafer)

was equipped with a 10X reduction capability, a 0.28-NA and *g*-line lens that could pattern a maximum square field of 10 mm x 10 mm. The lens was supplied by Carl Zeiss of Oberkochen, Germany. The stepper had a resolution of 1.25-μm lines and spaces over a 1.5-μm depth-of-focus and could achieve an overlay of ±0.7-μm. Its list price was about $300,000.

Several other companies soon began selling wafer steppers, and the major suppliers today are Nikon, ASM Lithography (ASML), Canon, and Ultratech. Modern steppers have considerably greater capability than the original GCA DSW4800, but the operating principles are essentially the same. Leading-edge systems in 2001 come in the step-and-scan configuration. (The stepper and step-and-scan configurations are described in more detail in Vol. 1, 2nd Ed., Chap. 13.) The characteristics of 248-nm step-and-exposure tools are listed in Table 7-1. To show how their imaging and overlay capabilities have been improved, these are listed both for the modern step-and-scan systems and the GSW4800 in that table.

7.2.1 Exposure Tools for 248-nm Lithography

Optical lithography systems continue to reach further and further beyond their previously perceived limits. By 2001 248-nm scanners – which use a KrF-laser light-source – were routinely printing 0.18-μm features. Industry experts believe that 248-nm lithography will get them to at least the 0.13-μm node. The pressure to push 248-nm lithography past what had been earlier considered its limits is largely a matter of economics. Chipmakers must weigh the trade-offs between the cost of the advanced reticles necessary to extend the current wavelength, and the systems costs and risks that come with the next wavelength. Manufacturers that run large volumes of devices per reticle are more willing to put effort into the current wavelength, spending the money needed on the reticles. On the other hand, those making logic or application-specific ICs (ASICs), where only a few hundred wafers may be manufactured with a reticle set, face a much higher cost of ownership if forced to pay the price of advanced reticles. Advances in resist technology and the ability to make lenses with higher numerical aperture are two of the reasons higher resolutions are possible with 248-nm lithography.[3] (Note that 248-nm resists are discussed in Chap. 6, Sect. 6.1.1.)

Lenses can now be built with higher numerical aperture (NA) and lower aberrations than was previously thought possible (Fig. 7-5). Even in 1999 it was not considered likely that lenses with an NA greater than 0.6 could be manufactured. Yet by 2001 various steppers were being introduced with even higher NA. For example, Nikon offers a stepper with a lens having an NA of 0.75, and the ASML PAS 5500/800 uses a lens with an NA of 0.8.

TABLE 7-2 Commercially Available KrF Step-and-Scan Systems - Compared to the First Commercially Available Stepper

	GCA	Canon	SVGL	Nikon	ASML
Model	DSW4800	ES3	Micrascan III+	NSR-S205C	5500/800
M:1	10×	4×	4×	4×	4×
Wavelength	436 nm	248 nm	248 nm	248 nm	248 nm
Maximum lens NA	0.28	0.73	0.60	0.75	0.80
Field size	10 mm × 10 mm	26 mm × 33 mm	26 mm × 34 mm	25 mm × 33 mm	26 mm × 33 mm
Overlay, Machine-to-itself	500 nm	25 nm	45 nm	30 nm	30 nm
Lens distortion	250 nm	20 nm	25 nm	20 nm	20 nm
Overlay, machine-to-machine	700 nm	45 nm	55 nm	45 nm	45 nm
Wafer size	3," 4," 5"	200, 300 mm	150, 200 mm	200, 300 mm	150, 200 mm
Throughput (200 mm wafers, except for GCA)	20 wafers per hour (4" wafers)	100 wafers per hour	90 wafers per hour	140 wafers per hour	120 wafers per hour

7.2.2 Exposure Tools for 193-nm Lithography

The ASML PAS 5500/900 ArF step-and-scan tool was introduced in 2001. It is based on the ASML 5000 architecture and uses a Carl Zeiss refractive Starlith 900® lens with CaF_2 elements for partial achromatization. The lens has an NA that is adjustable from 0.45 to 0.6. There is a large working distance from

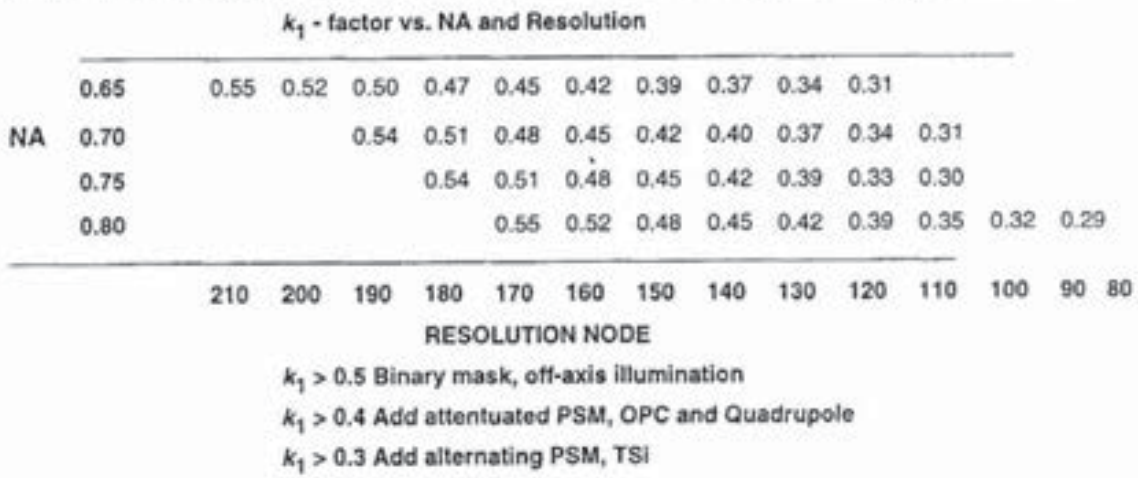


Fig. 7-5 The 248-nm lithography tools have seen vast improvements in lens aberrations in recent years. (Courtesy of ASML)

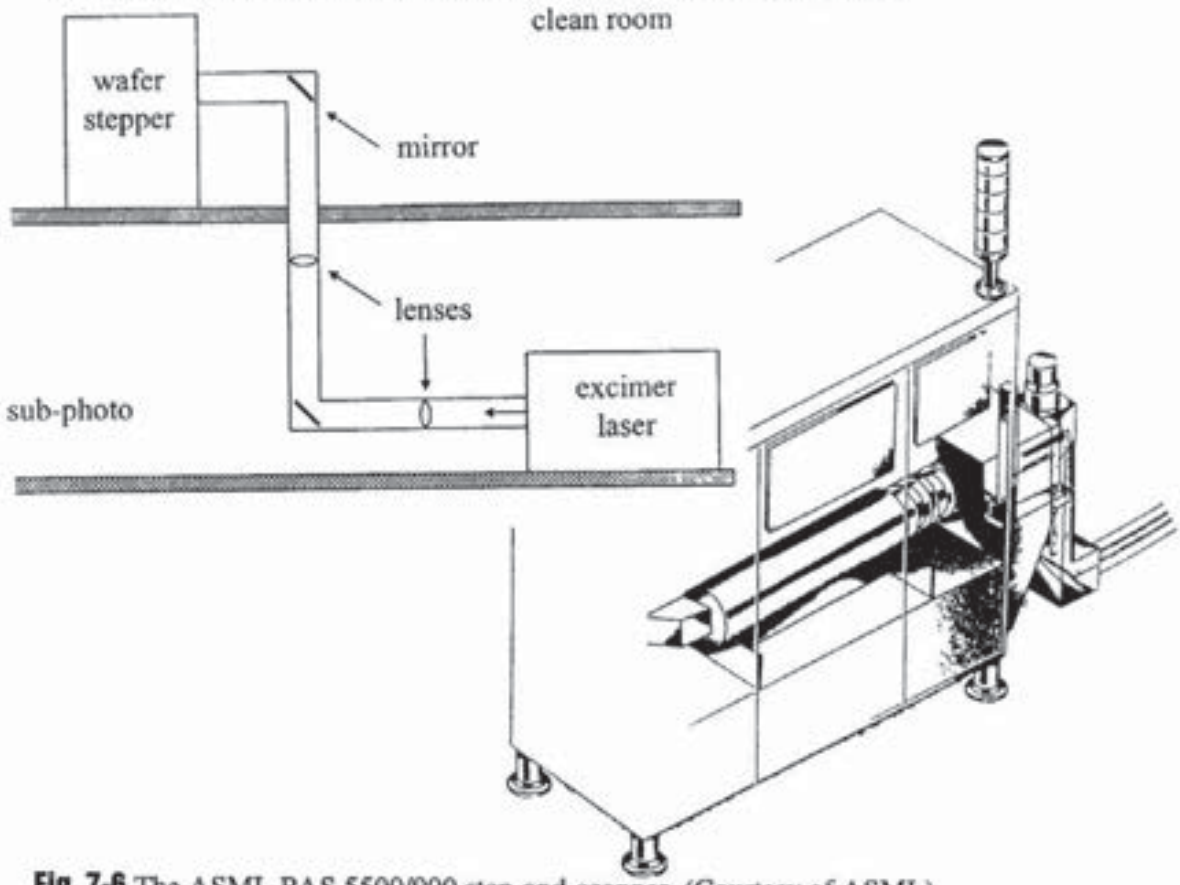


Fig. 7-6 The ASML PAS 5500/900 step-and-scanner. (Courtesy of ASML)

the final lens element to the photoresist surface, so that the lens contamination caused by outgassing from 193-nm photoresist during exposure is minimized. Figure 7-6 shows a picture of the PAS 5500/900 ArF step-and-scan system.

Canon's 193-nm scanner the FPA-5000-AS1 (introduced in 2000) has a 0.6-NA lens, and has demonstrated 600-nm DOF and can print 130-nm lines/spaces using annular illumination. The wafer throughput is projected to be 80-90 wph if the resist sensitivity is 10-mJ/cm^2. Challenges include availability of CaF_2 and improvements in resist performance.[4]

SVG offered two 193-nm exposure tools in 2000, the Micrascan V model (with an NA of 0.75) and Micrascan 193 (with an NA of 0.60). Both of these scanners use a catadioptric design, which requires only 1.7 kg of CaF_2 (as opposed to the 48 kg required for the chromatic correction of a refractive system).

7.2.3 Exposure Tools for 157-nm Lithography

Amid concerns about the technical risks and cost for all next-generation lithography (NGL) technology at and below the 70-nm node, work continues to develop 157-nm lithography technology.[5,77] This technology, however, is beset by a number of obstacles new to the evolution of lithography processes. First, the

material that has been used for the lenses and reticles of previous generations of lithography, namely fused-silica, will have to replaced by calcium fluoride (CaF_2). Second, the optical path will need to be a purged system requiring a nitrogen environment. Third, the optical system is likely to be of a catadioptric design, since such optical systems are smaller in overall design and require smaller optical elements than do comparable refractive optical systems. For example, due to the way the design is folded, the catadioptric design is approximately one third the size and weight of an equivalent all-refractive system, and the lens elements may be half that of all refractive optics in both size and numbers. While catadioptric lens elements could range from 50 to 200 mm, all-refractive lens elements as large as 400-mm might be required, depending on design.

The beamsplitter, the heart of the catadioptric optical systems, is a cube the size of which depends on the NA. The larger the NA the larger the beamsplitter that is needed. Fabricating beamsplitters from CaF_2 is a difficult challenge. As we will describe in the next section, the handling and polishing of this thermally sensitive and easily chipped material is a delicate operation. In 1999, however, SVG (in cooperation with Schott) announced that they have been able to successfully produce a 100-mm CaF_2 cube (Fig. 7-7).[6] Though not large enough for a 0.7-NA system, it meets the specifications for a 0.45-NA scanner.

At the First Symposium on 157-nm Lithography in 2000, ASML said it expects to ship its first 157-nm scanner in the 3rd or 4th quarter of 2003, and Canon expects to be ready about a year later. Estimates for the cost of a single 157-nm scanner range from $12m to $20M (for a system with a high NA).[7]



Fig. 7-7 SVGL has demonstrated the ability to grow and make a beam-splitter cube 100-mm on a side for use in its 157-nm catadioptric lithography tool. (Courtesy SVGL)

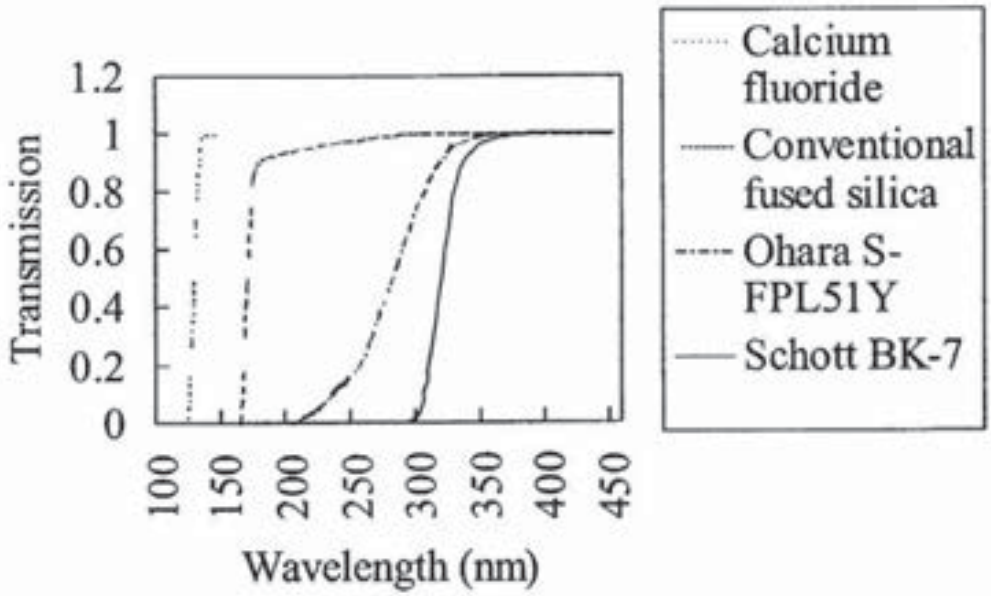


Fig. 7-8 Transmission of 10-mm-thick samples of materials used in lenses of wafer steppers.

7.2.4 Calcium Fluoride Optical Elements for DUV Lithography

Although the optical elements of lithography systems down to 248 nm can be made with fused silica, refractive optical systems at 193-nm often need to use at least some CaF_2 elements. At 157-nm, CaF_2 is virtually the only available useful transmissive material (with BaF_2 as the other possible candidate). The transmission properties of various types of optical glass are shown in Fig. 7-8. Lenses for 193-nm usually include CaF_2 as an optical material for some of their elements, especially at locations where the light intensity is especially strong. CaF_2 elements are used because damage is a concern under the high intensity levels of light delivered by excimer lasers. That is, one type of such damage is compaction of the lens material. Compaction occurring in the projection optics will degrade the image quality, and this compaction will also increase light absorption by the lens (and such increased absorption will therefore also reduce light intensity at the wafer, thereby reducing stepper throughput). Fused-silica undergoes compaction (becomes more dense) and thus also more absorbing under the intense light from excimer lasers. Consequently, CaF_2, which seemingly does not compact because it is a stable, single-crystal, is used in the illuminator and projection optics wherever high-levels of light intensity is expected. In 193-nm steppers, the compaction of fused-silica that occurs after extensive use is sufficiently severe so that it can render 193-nm lithography unsuitable for high-resolution lithography. In the illuminators of 193-nm systems, where such high intensities occur, appreciable amounts of CaF_2 are thus needed.

However, short wavelength applications also put very stringent requirements

on materials and fabrication. To minimize susceptibility to optical damage (as well as scattering and absorption losses that degrade throughput), bulk CaF_2 must be extremely pure and homogeneous. To reduce image degradation stemming from optical fabrication errors, the component surfaces must be flat and smooth to the level of tens or hundreds of angstroms rms. CaF_2 is hygroscopic, soft, and very prone to chipping, particularly at the edges. As a result, fabrication time CaF_2 for is typically 50% to 100% longer than for glass.

Besides transparency, there are other properties that a material must have in order to be used to make optical elements for DUV exposure tools. One such property is that the material does not exhibit excessive birefringence. (*Birefringence* is the optical property of a material in which a different index of refraction exists along various crystal axes.) However, many materials do exhibit significant birefringence, and such materials are generally not suitable for use in high-resolution optical systems. For example, SiO_2 in its purely crystalline form (quartz), is not easily incorporated into lenses because it has dissimilar behavior among different optical axes. On the other hand, amorphous fused-silica is a standard optical material because its non-crystalline structure causes it to have only a small degree of birefringence. Some materials, like CaF_2 have (until recently) been thought to exhibit little *intrinsic* birefringence. However, CaF_2 can exhibit low levels of birefringence induced by mechanical stresses. This type of *extrinsic* birefringence is termed *residual-stress birefringence.*[8]

Residual-stress birefringence occurs from the mechanical stress and strain formed during the production of an optical element (for example, if there are large thermal inhomogeneities in the furnaces in which the crystals [e.g., of CaF_2] are grown). A high level of residual-stress birefringence in optical components can lead to the aberration of light rays in an imaging system, and thus affect the imaging quality of a step-and-scan system. Residual-stress birefringence currently limits the yield of CaF_2 crystals suitable for use in stepper lenses.

Recently (May 2001), it was discovered that crystalline CaF_2 does, in fact, possess *intrinsic birefringence.* Lens designers were familiar with the residual-stress birefringence in CaF_2 described above, and were prepared to compensate for this effect with relatively complex lens designs. But, intrinsic birefringence – which may inflict about 6x more birefringence - presents a more sticky problem. However, a group of technologists gathered in mid-2001 to assess the issue, and they concluded that it will not be a showstopper for 157-nm lithography. Furthermore, intrinsic birefringence is not a concern for 193-nm lithography because corrections can compensate for any effects it causes at this wavelength. While it certainly will make 157-nm lens design more complex, the fact that this

phenomenon was discovered before such lenses had been fabricated is a good thing. It also implies that lithography systems will have to manage more carefully the polarization of the light as it enters a lens.[78] The lens design software (e.g., CODE V from Optical Research Associates) will also have to be modified to compensate for intrinsic birefringence.[9]

There are also questions about whether it will be possible to supply the quantity and quality of CaF_2 needed for 157-nm lithography. Nikon plans to use its own production for some of its optics supply. While, its major supply will not come from in-house, the company plans to develop the intellectual property for CaF_2 and optical coatings to help it get the quality needed. Corning Inc. recently acquired Optovac and Tropel Corp., to beef-up its capabilities to tackle CaF_2 production. It also has a facility in North Brookfield MA for producing CaF_2.

Note that while conventional fused-silica is not transmissive enough at 157-nm to be used as a 157-nm reticle-blank, its transmission can be improved by preparing the fused-silica with low-OH content (and by adding fluorine to the glass network). This increases the transmission to 82%, allowing fused silica to meet the International SEMATECH targets for 157-nm mask materials.[77]

7.2.5 300-mm Lithography Tools

Development of 300-mm tools began in about 1995. Before the end of 1997 the first 300-mm lithography stepper was installed (the Canon FPA-3000EX3L KrF-stepper) at I300I (now International SEMATECH). Initially, vendors were advised to develop tools that could handle only 300-mm wafers, but a few tool vendors took the bridge-tool approach (which allows a tool to be used on either 200-mm or 300-mm wafers). By late 1998, however, the industry was still not ready to embrace 300-mm production. Instead it stayed with technology-driven cost-reduction at 200-mm. By 2001, 300-mm had still not been widely adopted, as the industry went into a deep downturn. Nevertheless, mature 300-mm tools were ready, but the insertion node was predicted to be 130-nm, not at the 250-nm node, as originally envisioned.

It should be noted that stepper-scanners expose each field and then step to the next one. Since there are 2 1/4 times as many fields on a 300-mm wafer as on a 200-mm wafer, it could take twice as long to expose one. Achieving the desired throughput requires the wafer and reticle stages to move further and faster, while the overlay and resolution demands of 130-nm (and smaller) imaging reduce the tolerable errors and vibrations. All of the exposure tool manufacturers announced new products or platforms to meet these needs.[10,11]

ASML chose a "reaction-free-mass" concept for its new "Twinscan" platform, intended for operation both at 200-mm and 300-mm. To allow rapid acceleration of the relatively heavy wafer stages, the linear stage motors push against a countermass, which moves in the opposite direction, rather than against the fixed stepper frame. Since the countermass is free to move, it absorbs the impulse from the acceleration without inducing vibrations in the exposure tool structure. The center of gravity remains fixed, eliminating gravitational torques. The stages accelerate and decelerate according to Newton's third law, just like a rocket in vacuum (which spews reaction mass behind it). According to ASML, the new systems reduce moving-average alignment error to 3-nm and moving-standard-deviation to < 2 nm, a 10-fold improvement. The initial "Twinscan" tools were specified at 62-wph (in 2000), with throughput expected to double by 2002.

Canon chose a more conservative approach for 300-mm production, introducing a tool set that uses a high-NA KrF-scanner for critical levels, and a matching *i*-line-stepper for non-critical levels. All are designed for high throughput, specifically 88, 300-mm wph for the scanner, and 100 wph for the stepper. The matched fields allow for a mix-and-match strategy with the lowest CoO tool applied to each level. An ArF-tool is expected to be available in 2001. While Canon did not reveal its stage design, countermasses and improved aero-dynamics allow 350-mm/sec wafer speeds with overlay accuracy below 23-nm.

Nikon has had a 300-mm-capable KrF-scanner in the field since 1998 (the NSR-S203B), and the newly introduced 0.75-NA NSR-S205C promises 83 wph, along with 130-nm resolution (using RET) and < 30-nm overlay. Nikon expected to ship 100 of these systems in 2001. The NSR-SF100 *i*-line-scanner also matches the 25x33-mm field of the excimer scanners for mix-and-match applications. A new 0.68-NA ArF-scanner was promised for 2001.

Silicon Valley Group (acquired by ASML in 2001) introduced the Micrascan IV and V systems in 2000. These incorporate innovative low-mass precision wafer stages, without interferometer retroreflector mirrors. The interferometer systems of the Micrascan IV and V are inside out, compared to previous designs. That is, the mirrors are on the frame, and the laser-beam launching and collection optics are on the vertically oriented stages. The lightweight stages allow scan speeds above 300-mm/sec. The reticle stage, however, must move 4x-faster. The quick, four wavelength HPAS alignment system has 1σ overlay accuracy of 3-nm. The 193-nm Micrascan V tools are designed to expose 130-nm CDs with standard mask illumination and 100-nm half-pitch using RETs. These tools have a large NA (0.75) and catadioptric projection lenses. The 248-nm Micrascan IV with 0.60-NA optics is intended for non-critical applications.

7.2.6 Mix-and-Match Lithography

One way to lower the cost of lithography is to use the most advanced (and thus expensive) machines only for the critical levels, and employ less-expensive tools for the less-critical levels. This has been widely practiced for many years, and is referred to as *mix-and-match lithography.*

Vendors of exposure tools generally offer both high-resolution (expensive) and lower resolution (lower cost) models. The exposure tools intended for use on those levels with relatively lower-resolution requirements are normally optimized for throughput and price (instead of resolution and overlay). Lower-NA lenses are cheaper to design and build than higher-NA lenses. Such less-expensive systems deliver high throughput by exposing larger field sizes. In addition, in the 2001 timeframe all lower-cost models are steppers (i.e., not stepper-scanners).

Another strategy for lowering the capital cost of lithography is to utilize exposure tools from several generations of technology. As exposure tools get older they can no longer image the smallest feature sizes of the next technology nodes. Yet these steppers can still be used for imaging the less critical levels. This is an effective cost-saving method since it is not necessary to purchase new tools, and those in use may have already been fully depreciated. Smaller-NA steppers are also useful for patterning implant masks (that generally have larger feature sizes, but also often thicker resist films). In such cases, high-resolution steppers may actually do an inferior job, due to their limited depth-of-focus.

While mix-and-match lithography appears economically attractive, it is not without its drawbacks. It may involve using steppers from more than one supplier, or steppers that employ dissimilar operator interfaces. This necessitates extra training for the lithography workers, and a separate spare-parts inventory for each stepper model.

7.3 RESOLUTION ENHANCEMENT TECHNOLOGIES (RETs)

By 2001 it appears that the conventional lithographic processes that have enabled the descent of the IC feature sizes over the past 30 years, are approaching the limits of their usefulness. That is, even by using the shortest possible illumination wavelength, λ, and the largest numerical aperture (NA) lenses, the last factor, k_1, in the basic resolution equation:

$$\text{Minimum resolution} = k_1 \, (\lambda/\text{NA}) \qquad (7.1)$$

presents a formidable obstacle. (Note that k_1 represents the measure of lithographic aggressiveness.)

That is, to obtain acceptable yield with a conventional lithography process (i.e., one which does not use any resolution enhancing techniques), the value of

k_1 is restricted to values larger than ~0.75 (because unacceptable image degradation occurs below this value). However, in the past two decades, various *resolution enhancement techniques* (*RETs*) have been developed, and their use permits lower values of k_1 to be achieved.[12,13] These RETs include the use of anti-reflective coatings and top-surface imaging (as is discussed in Chap. 6), and off-axis-illumination (OAI), optical-proximity-correction (OPC), and phase-shift-masking (PSM), which are discussed in this section. Use of these techniques can possibly reduce k_1 from 0.75 to 0.35, which represents more than a 50% decrease in critical dimension compared to conventional lithography. Thus, we have gained the capability to print features that are one-half the size of those that can be printed with conventional lithography. Hence, by employing RETs, optical lithography has been able to maintain its vitality - allowing it to be used much longer than was ever thought possible 10 years ago.

7.3.1 Off-Axis-Illumination

Off-axis-illumination is a method for improving the resolution of a stepper without resorting to a higher NA lens or a shorter wavelength of light. The method was long employed in optical microscopes to enhance image contrast, and was introduced to microlithographers by Mack[14] and Fehrs[15] in 1989. Researchers at Canon, Inc.,[16] and Nikon Corporation[17] incorporated it into the optical systems of steppers in 1992. The off-axis-illumination technique improves the resolution capability of the stepper by modifying the direction of incidence of the illumination on the reticle. That is, it directs the beam of light from the illumination system of the stepper through the reticle in a direction that would make it strike the projection lens at the edge of the entrance pupil (rather than at the center of the projection lens entrance pupil, as is done in the conventional illumination technique). The term *off-axis-illumination* is somewhat of a misnomer. This is because conventional partially-coherent illumination also includes a range of angles of incidence on the mask, both on-axis and off-axis. Thus, when the term "off-axis-illumination" is used to describe the *post-1992 techniques*, it is intended to mean an illumination scheme *without any on-axis components*.

The principle of the new "off-axis illumination" techniques can be understood by considering the imaging process of optical lithography from a Fourier optics perspective (see, for example, Vol. 1, 2nd Ed., Chap. 13), as applied to imaging the pattern of a grating of lines and spaces. As shown in Fig. 7-9, in conventional stepper optical systems the light beam is directed from the illumination system through the reticle in a direction which would coincide with the center of the entrance pupil of the projection lens. Recall that for a grating pattern, the light is

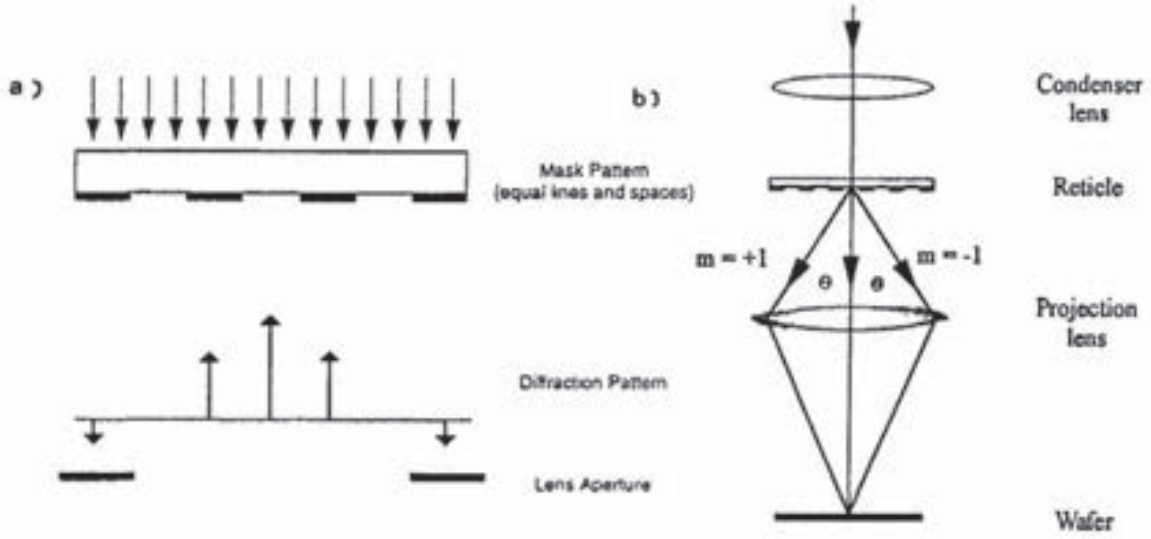


Fig. 7-9 Principle of on-axis illumination: a) Light illuminating the mask from an on-axis direction, and the diffraction pattern from an on-axis coherently illuminated grating mask. The pattern is centered on the entrance pupil of the objective lens. b) Aperture with on-axis illumination showing the 0-order (undiffracted beam) and the two first order diffracted beams, m = +1 and m = -1.

diffracted into a set of beams which propagate away from the grating only at specific angles (corresponding to the orders of the diffracted beam). The undiffracted beam (0-order) propagates at a direction perpendicular to the reticle plane. The first-order diffracted beams (m = +1 and m = -1) propagate at angles $\theta_1 = \sin^{-1}(\lambda/p)$ and $\theta_1 = -\sin^{-1}(\lambda/p)$, respectively (see Fig. 7-9) where p is the pitch of the grating (i.e., the distance equal to the sum of the line and the space). The third-order diffracted beams propagate at larger angles than the first-order beams (i.e., $\theta_3 = \sin^{-1}[3\lambda/p]$ and $\theta_3 = -\sin^{-1}[3\lambda/p]$). Since no imaging is possible unless at least one additional beam is accepted by the lens besides the 0-order beam, the NA of the lens must be large enough to accept both the +1 and the −1 first order beams. For a given NA and λ, there is a minimum pitch below which the first order beams will fall outside the collection area of the lens. For this pitch and smaller, no imaging of the grating pattern is possible. An optical system that collects the three beams (0-order and +1 and -1 order beams) is sometimes called a system that employs *three-beam imaging*. Note that if at least the first order beams are captured within the pupil of the lens (and for the case of coherent illumination), this leads to the simple resolution limit of 0.5 λ/NA, or in equivalent terms, a k_1-limit of 0.5. For gratings with pitches below this limit, no pattern is formed by an optical system with this λ and NA, using coherent light.

If an *off-axis* (or *oblique*) illumination scheme is employed (Fig. 7-10a), such

that the illuminating beam is directed so the 0-order beam strikes the entrance pupil of the projection lens just at its edge, and only one of the other first-order beams strikes the other edge of the entrance pupil (i.e., either the m = +1 or the m = −1 beam) only two-beams are collected (hence the name *two-beam imaging,* see Fig. 7-10b). In this case the angle, θ, can have a value greater than in three-beam imaging, implying that the pitch of the grating can be smaller (and in turn a higher resolution can be achieved for an optical system employing the same NA and λ). In the extreme case of this type of off-axis illumination, a doubling of the resolution-limit capability of the imaging system is obtained, compared to a coherent fully on-axis-illumination scheme. Such resolution-enhancement is most pronounced for the case of dense-patterns of equal lines and spaces (such as occur in the regular array patterns of DRAM and SRAM interconnect lines).

Although resolution-enhancement may seem to be the primary benefit of off-axis illumination, in fact it is the impact of illumination on DOF that provides the more significant advantages. Defocus causes a phase error for each diffraction order that is proportional to the square of the radial position within the pupil. Consider the case of off-axis-illumination where the zero-order and one of the first-orders are the same distance from the center of the objective lens pupil. In this case, both the zero-order and the collected first-order will have phase errors due to defocus, but these errors will both be the same, and the relative phase difference between the zero- and first-order beams will thus be zero. Such a circumstance will result in an image that is much less prone to defocus, resulting

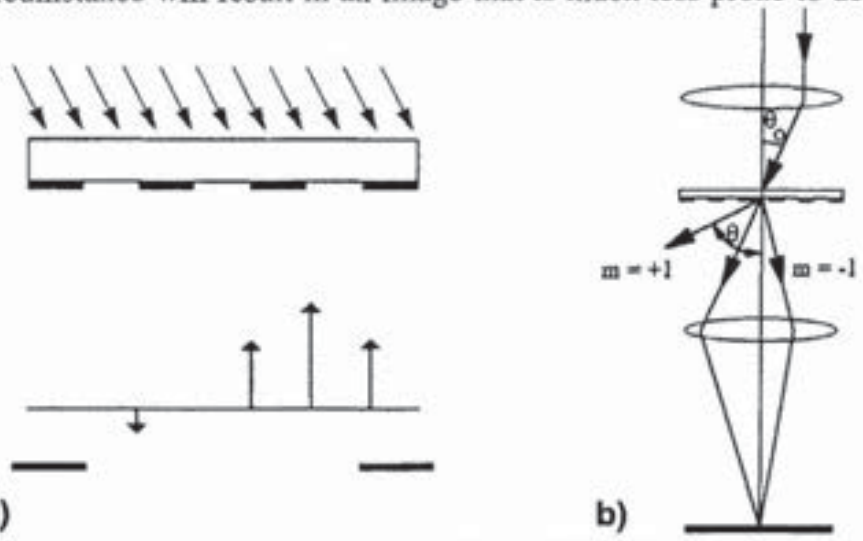


Fig. 7-10 Principle of off-axis illumination: a) Light illuminating the mask from an off-axis direction, and the diffraction pattern from an off-axis coherently illuminated grating mask is no longer centered on the entrance pupil of the objective lens, but is instead shifted. b) Aperture with one off-center hole that produces the illumination shown in (a).

in an important improvement in DOF. Thus, the proper choice of an off-axis illumination angle for a given equal line/space pitch will result in a dramatic increase in the DOF of that feature.

One limitation of off-axis illumination is that it does not provide the same resolution enhancement for isolated lines as it does for dense lines and spaces. This is because discrete diffraction-orders do not exist for isolated features. Instead an isolated feature generates a continuous diffraction pattern. Hence, isolated line performance is only minimally improved by off-axis illumination. When both dense and isolated features are considered together in a field, a dense-to-isolated feature-bias will arise if off-axis-illumination is employed.[18] To circumvent this problem, off-axis-illumination tricks have been proposed. One involves adding small (sub-resolution) auxiliary patterns close to an isolated feature to establish the required interference effects (see Sect. 7.3.1).[19,20,21] Another issue that must be addressed is that some illumination energy is lost because one of the 1st-order beams is not collected. As a result, it is necessary to increase the exposure time to compensate for this loss of illumination.

If the illumination of the mask only comes from one beam (that is, through an aperture with only one off-center hole, as shown in Fig. 7-10b), the +1st-order will fall completely out of the pupil, but the 0th and -1st-orders will be collected to form an image (Fig. 7-10b). However, the asymmetry between the 0th and -1st orders leads to severe telecentricity errors. Symmetry can be restored by illuminating with two sources, placed on opposite sides of the pupil, using an off-axis illumination aperture as shown in Fig. 7-11a (*dipole illumination*). This is an improvement over the single-beam aperture, but it will only improve imaging for gratings oriented in one direction. It will degrade them for gratings oriented in the perpendicular direction (see Fig. 7-11a1). Nevertheless, if geometries are oriented in only the vertical direction, such dipole illumination is optimal. *Quadrupole illumination* (Fig. 7-11b) works well when all features are oriented left-right or up-down, but distorts features of all other orientations (e.g., a line at a 45° angle). The latter limitation can often be tolerated because critical features are rarely oriented at odd angles to the sides of the chip. The *annular aperture* shown in Fig. 7-11c improves features of all orientations, but at the price of milder improvements in the resolution and DOF.

Note that a specific off-axis illumination setting will improve imaging performance for only a narrow range of pitches, but this is still an advantage for many applications, such as memories, where many structures are highly periodic.

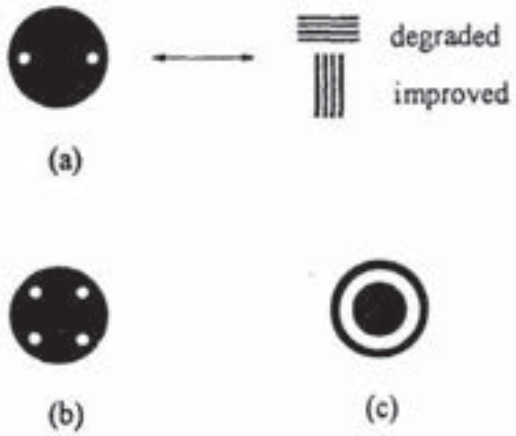


Fig. 7-11 Different types of off-axis illumination apertures. The aperture with holes at the edges or corners - (a) and (b) – will improve lines oriented in one direction (see a1), but will degrade imaging of lines in the orthogonal direction. (c) Annular off-axis illumination aperture.[23]

However, an off-axis illumination setting optimized for one pitch may work poorly for others. For patterns with more than one pitch, this becomes a problem. In such cases, an off-axis-illumination source that includes some on-axis illumination (with lower intensity than off-axis light), reportedly gives better results.[22]

Off-axis-illumination is relatively easily implemented. Stepper manufacturers can provide any of the illumination-aperture types described above. A series of these apertures can even be mounted on a turret, and the particular aperture desired for any mask can be automatically supplied by the stepper control program. More sophisticated illuminators can also be used to transmit a larger fraction of the available light into the annulus of illumination.

7.3.2 Optical Proximity Correction (OPC)

The techniques developed to extend optical lithography beyond the range of conventional imaging are based on the idea that the 2-D aerial image formed by the projection lens is a structure to be optimized. One way to optimize the aerial image is to pre-compensate the reticle pattern to account for expected pattern distortion due to diffraction effects. One important consequence of diffraction is the so-called *proximity effect*, in which the exact size and shape of the features patterned on the wafer depend on their proximity to other features. These proximity effects manifest themselves in the form of distorted geometries (and in some extreme cases, even unprinted patterns). For designers, this phenomenon translates into line-width variations and other distortions that can significantly

decrease performance – or even worse - cause missing, incomplete, or shorted structures that can result in circuit failure.

One technique for determining such expected pattern-distortion involves computer simulation. That is, the 2-D aerial image of a particular reticle feature (or a group of features) is simulated by calculating the effects of diffraction on the image.[24] Based on the results, the reticle pattern can be altered and simulated again to determine if the change has improved the 2-D aerial image.[25] The reticle patterns are altered by adding (or subtracting) sub-features to the original pattern, and these features are called *optical-proximity-correction-features*. Additional aspects of the optical-proximity-correction procedure is discussed next.

One effect long observed in microlithography is that when patterns smaller than the exposure wavelength are being imaged, individual pattern features do not image independently. Instead, they interact with neighboring patterns. A simple qualitative argument for why this occurs is based on the light intensity pattern of the Airy function. That is, overlap of the secondary diffraction peaks of the Airy function with adjacent pattern features leads to increased or decreased exposure intensity at any point in the image compared to a purely geometrical image model. For example, the linewidth of isolated- and densely-grouped-lines eventually varies by a significant amount from that of the predicted linewidth (based on the pattern size on the reticle) when the feature sizes approach the

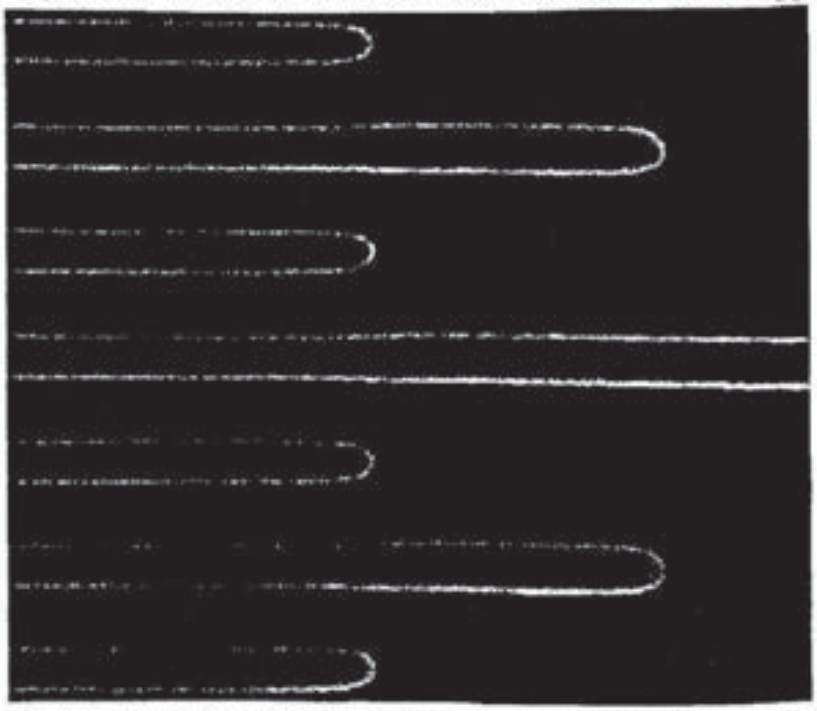


Fig. 7-12 Optical proximity effect. All pattern features are the same size on the reticle, but the final image is larger for the isolated line extension than for the densely packed lines.

wavelength of the illuminating light. However, very little difference between two such aerial images is observed if the feature sizes are large, i.e., > 1 μm. For example, assuming the use of positive photoresist, 0.35-μm-wide *isolated* lines are 0.09 μm wider than *closely spaced* 0.35-μm-wide lines (i.e., which might be part of a diffraction-grating pattern). Figure 7-12 shows an example of this effect. Since the linewidth depends on the proximity of other features, this effect is called an *optical-proximity-effect.* Such proximity-dependent linewidth variation (specifically termed *iso-dense print-bias*) may become significant in the interconnect-lines of a memory chip. That is, lines within the memory-array region of the chip would be regularly and closely spaced, but in the peripheral sense-amplifier and logic-circuitry they could be spaced farther apart. If this dimensional-difference can be predicted, the width of the chrome lines on a reticle can be adjusted to make isolated and densely-spaced printed lines on the wafer all the same width.[26] Such adjustments to the sizes of patterns on a reticle are termed *optical-proximity-corrections* (*OPC*), and in this case such an OPC is performed as a *one-dimensional, pattern-dependent width correction* to lines on a reticle. (Note that biasing the size of features has been used historically for many years to assist in lithography-process centering and optimization, but until recently the same bias was applied to the entire reticle. On the other hand, the OPC-bias is custom-tailored for each line in its own environment.)

Originally, OPC was considered to be just a method to correct the line-size variations as a function of the proximity of other features (as described above).[27] However, other pattern-modification techniques have been developed to improve the images of such two-dimensional IC features as contact holes (or vias) and lines (e.g., of poly or metal). That is, corner-rounding and general loss of shape fidelity in small features is caused by the inability of the projection lens to

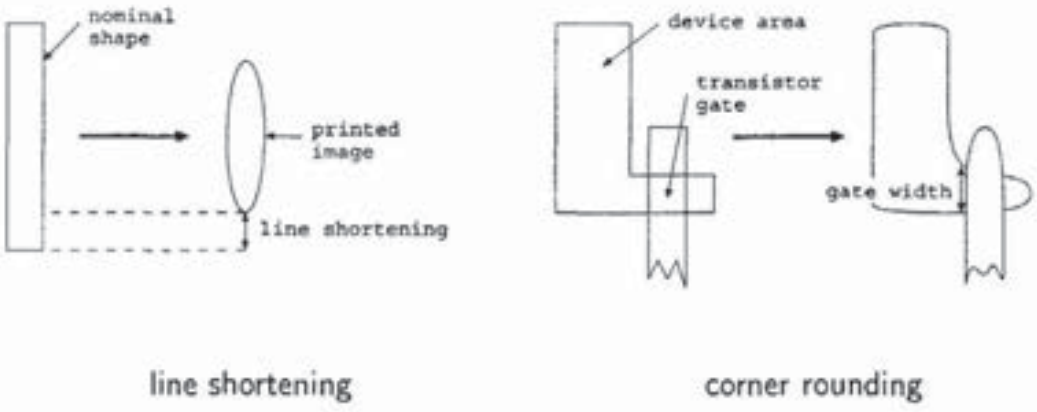


Fig. 7-13 Example of line shortening due to image distortion.

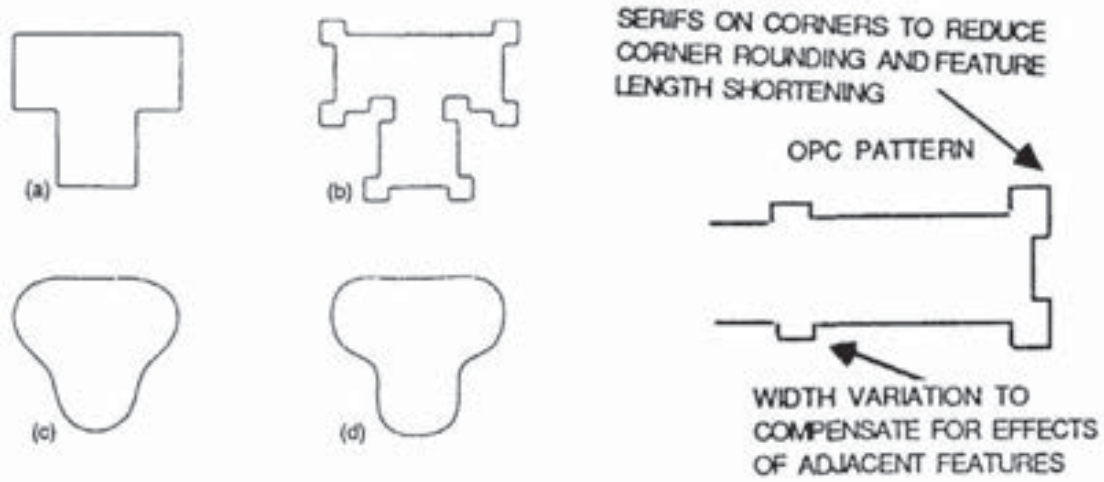


Fig. 7-14 A T-shaped feature (a) with dimensions near the resolution limit of the projection lens is printed as a rather featureless blob (c) on the wafer. Addition of serifs (b) brings the printed image (d) closer to the shape originally designed. Reprinted by permission of Marcel Dekker. (e) Adjustments to the linewidth can correct for the proximity of nearby features, and serifs on the end of a line can correct line shortening effects.

resolve details smaller than the diffraction limit of the lens. For example, a square contact-hole pattern with a size close to the minimum resolution limit of the optics will print more nearly like a circle. Likewise, a line (which should print with a rectangular shape - see Fig. 7-13a), instead prints as a feature with an elliptical shape (and an even shorter length). These effects have therefore also been classified as another form of optical proximity effect.

The OPC techniques used for modifying the 2-D reticle patterns to alleviate such deviations in the aerial images involve appending small, secondary patterns to the primary contact hole or short poly- and metal-line patterns. As noted above, such added-on auxiliary features are intended to be below the resolution capability of the projection optics. Several types of OPC features are employed. The most common are secondary patterns (called *serifs*) which are added to the corners of square hole patterns (as shown in Fig. 7-14b & e). They make the aerial image (and hence also the printed image) of such small contacts become more square-like. This occurs because the serifs enhance the amount of light transmitted though the corners of transparent mask features (e.g., the pattern of a contact hole being imaged in a positive resist). For short lines, serifs (together with linewidth changes) may be also added to improve the fidelity of the printed feature. The second type of OPC feature is the *line-jog*. These are used for lines that are isolated along part of their run across the wafer surface, but are near

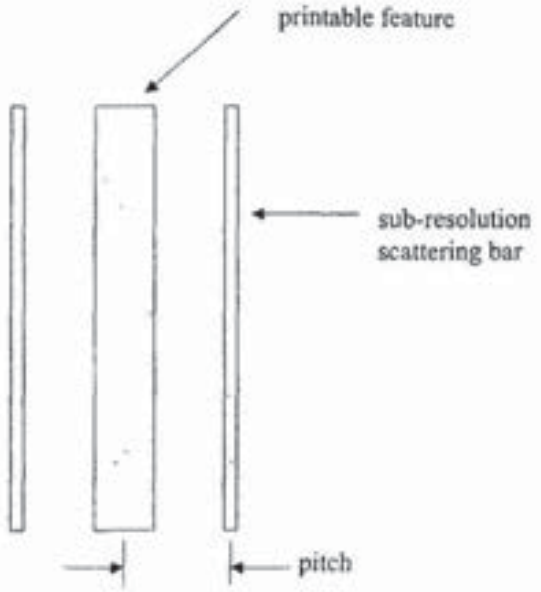


Fig. 7-15 Sub-resolution scattering bars applied to isolated features.

other lines at other places. For such lines, incremental linewidth changes can be made to the linewidth by employing such line-jogs (see Fig. 7-14e).

A third type of OPC feature is the *scattering bar,* whose role is to improve the imaging of isolated lines. Scattering bars are lines having a sub-resolution width that are placed adjacent to isolated lines on the same pitch as dense features on the same reticle (Fig. 7-15). The purpose of scattering bars is to create a dense environment around an isolated feature. However, when properly designed, these scattering bars do not print. The isolated line thus appears dense to the projection system, but prints as a fine, isolated line.*

While OPC techniques provide significant benefits, they are not gained without some costs.[18] The adoption of OPC includes the following drawbacks. First, serifs add a large number of additional features to the database of the reticle pattern. This increases the time needed to write the reticle (and therefore also its cost). Second, the sizes of the serifs are typically much smaller than the minimum feature size of the reticle, which creates problems during the inspection phase of the reticle manufacturing process. Likewise, when sub-resolution features first appeared, they represented a decrease of minimum feature sizes by factors of three and four. Such large changes in minimum feature size necessitated rapid advances in mask patterning. Third, this technique may not be feasible in all situations. That is, if the spacing between two contacts is already at

* As noted earlier, off-axis-illumination is an effective method used to enhance the patterning of periodic structures, but it provides no benefit for isolated lines.

the minimum allowed dimension, adding serifs to the square contact-hole pattern would bring parts of the two contact hole patterns to a distance below the minimally allowed value, and in such cases they cannot be used. Finally, it must be remembered that OPC is intended to correct for proximity effects. It does not actually improve the resolution or other quality metrics of the lithography process.

Since OPC is not performed until the design data are prepared for mask fabrication, OPC is typically transparent to circuit designers (although in some cases the design team may elect to get involved in deploying OPC in their layouts). However, in the first case, designers must still select a chip-manufacturer and mask-maker capable of accommodating OPC adjustments to the design. In both cases, some method must nevertheless be employed to determine which types (and where) OPC features have to be added to circuit layouts. Two approaches have been developed for carrying out this latter task.

The first is the so-called *rules-based* method. This uses a pre-determined set of rules together with a pattern-layout database program. The database program systematically determines the proximity environment of every pattern feature-edge, and then uses a lookup-table algorithm to apply the appropriate rules. These rules can generate biasing, serifs, scattering bars, etc. The rules may be developed directly from experience in the wafer fab, or they may be determined from image and resist simulations. Such rules-based approaches to determining the addition of OPC features to mask patterns are generally quite fast (requiring from several minutes to several hours on a typical workstation, per pattern level). However, the process of generating and validating the rules themselves can be a lengthy and tedious process. Since the accuracy and precision of the OPC features are completely determined by the extent and quality of the rules, this task must be effectively performed.

The second method is the so-called *model-based OPC approach.* In this case, each feature and its proximity environment are modeled by an image analysis program. A set of mathematical model parameters is used within the program to determine the appropriate corrections to the pattern layout. The model-based method is computationally intensive and may require many hours (or even days) of processing per pattern level. The model parameters are very important, but there are a limited number required. Manufacturers fabricate special test chips to generate OPC models of their lithography process. The chief advantages of the model-based approach is that the correction accuracy and precision can be readily controlled, and that corrections can be successfully applied to any sort of pattern layout. Software companies that offer OPC programs include Mentor

Graphics, Numerical Technologies (NumeriTech), and MaskTools (an ASML Company).

7.3.3 Phase-Shift Masking (PSM)

The most dramatic improvement in resolution limits imposed by diffraction involves the use of the phenomenon of *phase shifting*. For example, if a lithography process can attain a resolution of 0.3-μm using a 248-nm KrF light source with a stepper having an NA = 0.6, then by incorporating phase-shifting techniques on the reticle, the resolution can be improved to as much as 0.18-μm. The concept of increasing the resolution of a lithographic image by modifying the optical-phase of the mask transmission was first suggested in 1982 by Levenson, Viswanathan, and Simpson.[29] While a number of different types of phase-shifting masks have subsequently been developed, they all employ the same basic concept, as illustrated by the original version. This concept will be used here to demonstrate the principle of *phase-shift masking* (PSM). Several references review the topic of PSMs in more detail than is provided here.[30,31]

The phase-shifting of light as it passes through the clear part of a mask is accomplished by adding an extra layer of transmissive material to the optical path (Fig. 7-16). For PSM applications, a phase shift of 180° is usually desired, and this corresponds to an optical path-length difference of $\lambda/2$ at the stepper wavelength. Assume now that a grating (representing a *binary imaging mask*, or *BIM*, Fig. 7-17a1) is being imaged by optics with a particular numerical aperture

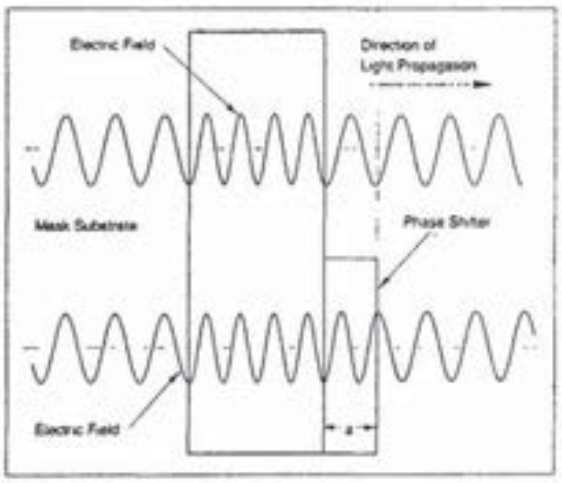


Fig. 7-16 The phase-shifting layer introduces a difference in optical-path-length. Here the difference is $\lambda/2$.[30] © IEEE (1993)

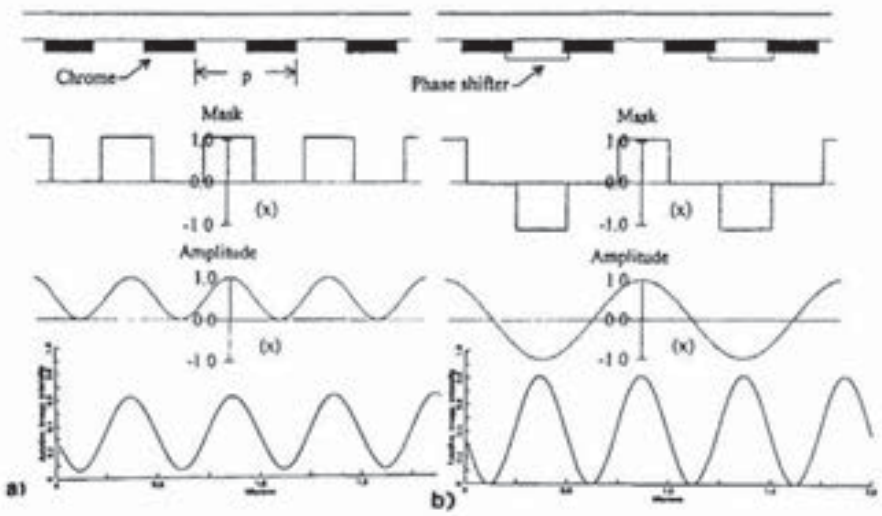


Fig. 7-17 Schematic of (a) a conventional binary mask (b) an alternating phase shift mask, (ALT PSM). The mask electric field, amplitude, and image intensity is shown for each type of mask.[28] Reprinted with permission of Marcel Dekker.

and partially coherent light. The (normalized) amplitude of the electric field (*E*-field) at the mask is either +1 or 0, as shown in Fig. 7-17a2. However, the imaging is degraded at the wafer because light from the clear areas on the mask is diffracted so that it strikes regions on the wafer that would ideally be completely dark. The nominally-dark-regions have light diffracted onto them from the clear spaces on both the right and left of them. The *E*-field amplitude on the wafer as a result of such diffraction is given in Fig. 7-17a3, and the light-intensity (which is the square of the *E*-field-amplitude) is plotted on 7-17a4.

If a phase-shifting material is placed on *alternating* clear spaces of a mask (as shown in Fig. 7-17b1), this forms a type of PSM called an *alternating-PSM*, or *ALT-PSM* (or sometimes even a *Levenson-PSM*). It should be noted that producing a phase shifter on a clear region of a mask may be done in one of two ways. In the first, an appropriate film of material can be deposited over the entire mask, and then etched away in the desired areas. In the second, the fused-silica glass itself can be etched in the appropriate clear-areas to produce recesses in the glass. The latter technique requires well-controlled etches so that recesses of the appropriate depth can be produced with only a timed etch. However, since deposition of a film different than fused-silica is not necessary, the use of this second approach to produce ALT-PSMs avoids the potential problem of the phase-shifter affecting either the light amplitude or phase.

In an ALT-PSM the light is phase-shifted by 180° when it passes through the spaces containing the PSM material, and the *E*-field amplitude on the mask is

shown in Fig. 7-17b2. Note that the *E*-field amplitude at the mask spaces containing the phase shifters is -1, but at those without a phase shifter has a value of +1. The *E*-field intensity imaged on the wafer using such a PSM is shown in Fig. 7-17b3, and the corresponding light intensity on Fig. 7-17b4. It can be seen that the contrast in the light-intensity distribution of the image on the wafer is greater when the PSM is used. Contrast is increased because the light diffracted into the nominally dark area from the left-clear-area will interfere destructively with the light diffracted from the right-clear-area. All types of PSMs use this same principle, in which the destructive interference of light of opposite phases is used to improve image contrast.

While ALT-PSMs give the greatest improvement in resolution for the case when closely spaced clear lines are separated on an opaque background, this type of mask is not effective for some other types of patterns. One example is the case when tightly packed features are layed-out in a pattern that resembles the stacking of bricks in a wall (Fig. 7-18). Since such a pattern has alternating rows offset from one another, it is not possible to assign a phase value to each pattern that will make it 180° out of phase with all its neighboring features. The resolution of isolated features (such as contact holes) is also not improved by ALT-PSMs. In general, non-repetitive patterns cannot be given phase assignments that meet the ALT-PSM approach.

Another problem with ALT-PSMs occurs when opaque lines are created on a clear background. At the edge where the clear area covered with a phase shifter abuts an area of clear-glass without a phase-shifter, interference between the two phases will make such boundaries print as a narrow dark line (Fig. 7-19a). A possible cure for this problem is to put another layer of material at such boundaries which shifts the phase by only 90° (see Fig. 7-19d), but in practice such a cure is not completely satisfactory. That is, it requires additional steps in an already complex reticle-fabrication process. Another approach to solve this problem involves the use of negative photoresist to produce masks for IC levels that produce line patterns. By using negative resist, if patterns on a mask which define *lines* are clear-features (and those defining spaces and field-regions are opaque, Fig. 7-20), phase-shifter material never abuts a clear area. This is one of the reasons that negative resists are again being considered for deep-submicron lithography applications that call for PSMs. A lower-cost version of ALT-PSMs (the *sidewall-chrome alternating aperture* [*SCAA*] PSM) specifically tailored for small production runs (< 500 wafers), was recently described by Levenson.[32] In this type of PSM, the phase-shifting regions are patterned into the mask substrate before the chrome is deposited onto the mask blank. Then the chrome is pattern-

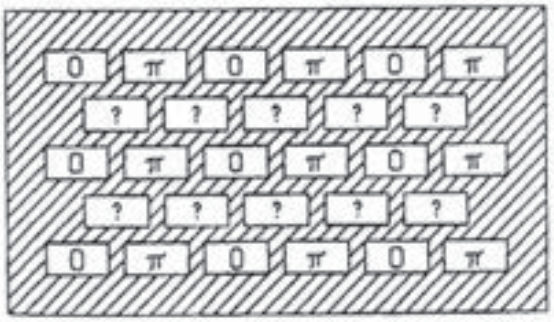


Fig. 7-18 Limitation of ALT-PSMs. There is no way to assign phases so that a 180°-phase difference occurs between adjacent transparent features In this example, the phases have been assigned to the odd-numbered rows, but there is no way to assign phases consistently to the even rows.[36] Reprinted with permission of Marcel Dekker.

ed with a second exposure. To make the masks less costly, about 1000 mask blanks would have to be layed out with pre-determined phase-shift locations. Special EDA tools would then be used to lay out chips so all that all fine-features lie at these predetermined phase-shift locations. Levenson opines that this situation would be compatible with >90% of ASIC designs.

Other types of PSMs have also been developed in an attempt to overcome the drawbacks of ALT-PSMs for some applications. Before describing these other types of PSMs it should be mentioned that the various types of PSMs are grouped into so-called *strong* and *weak* PSMs. Therefore, these terms must be defined. The distinction is based on the maximum degree of improved resolution and depth of focus that each type can provide. Strong-PSMs (such as ALT-PSMs) can provide up to a 40% improvement in resolution, but a lesser improvement in depth of focus. Weak-PSMs (such as the *rim-shift-PSMs* and the *attenuated* [or *halftone*] *PSMs* which are described next) offer less improvement in resolution, but more improvement in depth of focus. The benefits each group offers are different, and this implies that the "best" type of PSM depends on the particular application (as described in more detail later).

One of such alternative type of PSMs is the *rim-shift-PSM*, which falls into the group of weak-PSMs. The rim-shift-PSM was developed to address the problem of overlay in PSM structures by creating a self-aligned phase-shifting structure. In the fabrication of such structures, a phase-shifter film is deposited onto the chrome layer and then resist is spun onto the reticle. The resist is then patterned by e-beam exposure and developing.

This is followed by an anisotropic etching of the phase-shifter film. Finally, the chrome film is etched isotropically so that it undercuts the phase-shifting film

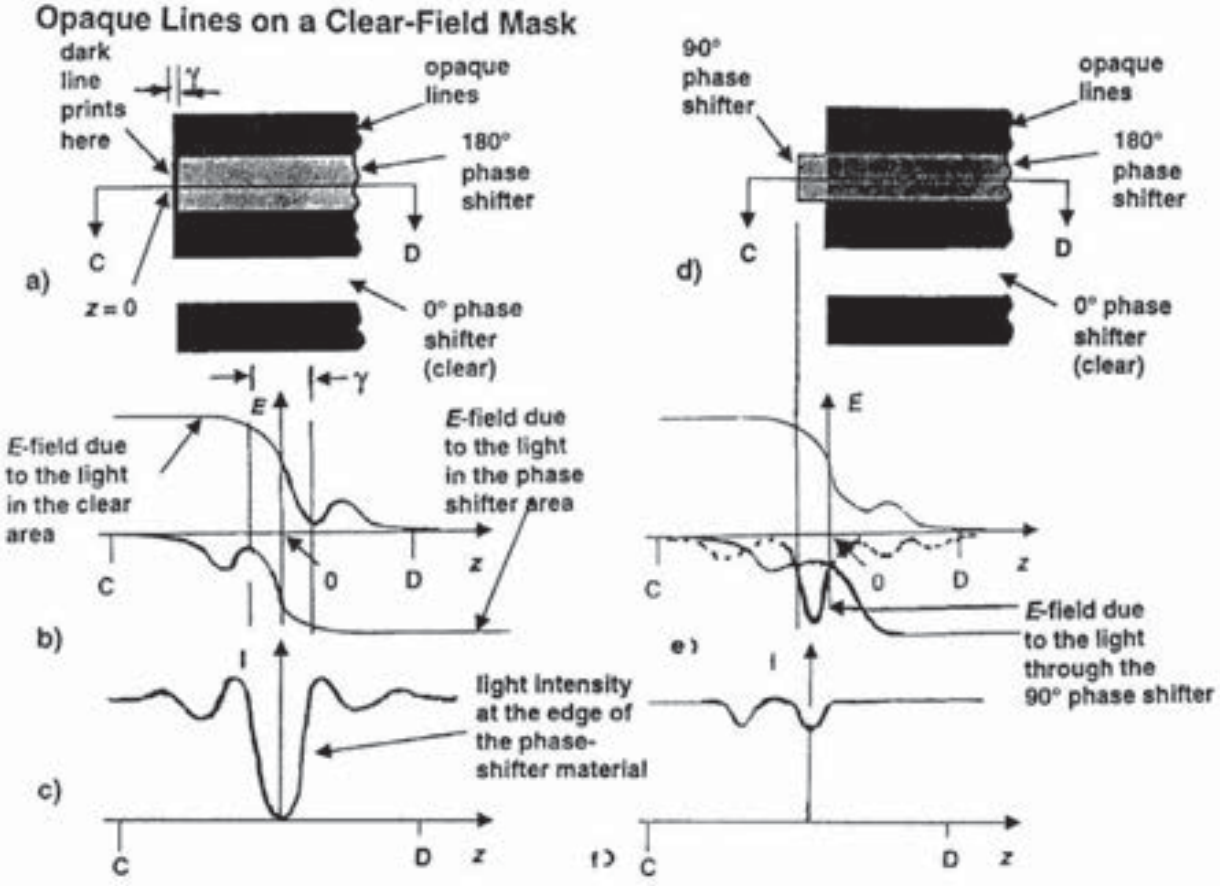


Fig. 7-19 Problem at the edge of a phase-shifter at a clear boundary: (a) ALT-PSM used on lines and spaces pattern with a clear field. Dark lines will print at the boundaries between the edge of the phase shifter and non-shifted clear areas; (b) *E*-fields along the *z*-axis between C and D for light incident on the phase shifter and clear area; (c) These two *E*-fields interfere in the vicinity of $z = 0$, causing the intensity of the *E*-field there to drop, resulting in a dark line printed as shown in (a); (d) Adding an extension of phase shifting material (with $\phi = 90°$) at the end of the phase-shifting region (as shown in [d]), helps mitigate the problem; (e) The amplitude of the *E*-fields of the light for the three phase regions versus *z*; (f) Intensity of light versus *z* when the $\phi = 90°$ extension is present. Note that the intensity drop-off is much smaller when interference occurs, and the dark line may no longer print.

(see Fig. 7-21). In rim-shift -PSMs the phase -shifting takes place only at the rim of each mask feature. As a result, such PSMs do not provide as much resolution improvement as do the ALT-PSMs. However, they do not suffer from the pattern restrictions that hinder ALT-PSMs. Thus, they can be used for arbit-

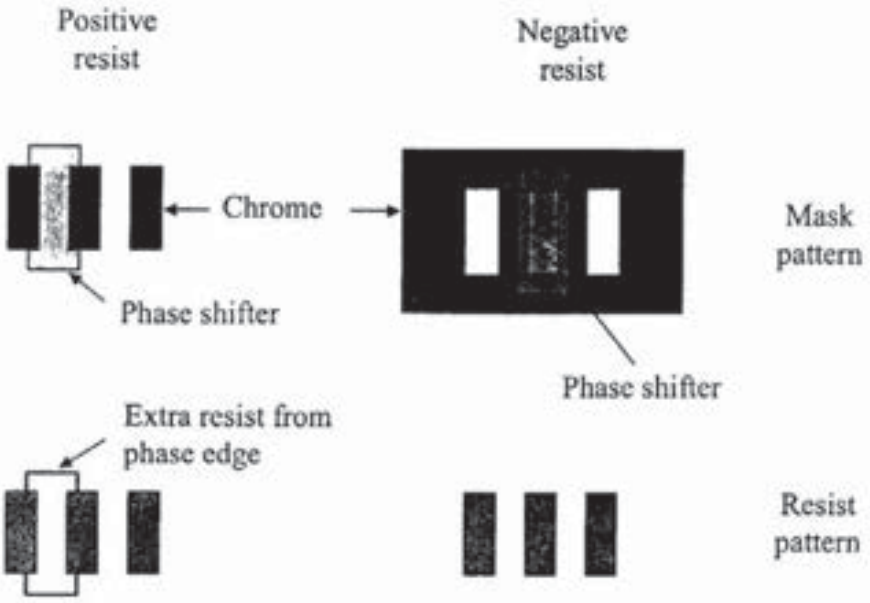


Fig. 7-20 On the left, extra resist printed by the phase edge, for positive photoresist. The right side of the figure shows what occurs with negative resist.

rary reticle patterns, including isolated features. Among the issues of concern are rim-size optimization and the lack of inspection and repair of the phase-shifting materials of the mask.

Another type of weak-PSM that can also be applied to arbitrary layout patterns is the *attenuated* (or *halftone*) *PSM*.[33,34] Such *attPSMs* are fabricated by replacing the opaque part of a conventional reticle with a halftone film (i.e., one that is partially transmissive, about 10%). For example, for *i*-line-light such a film can be made from a very thin layer of chromium oxynitride or Mo-Si-ON (molybdenum silicon-oxynitride) having the proper thickness to shift the phase of the light it transmits by 180° (Fig. 7-22). While some light will pass through the halftone film, it will be too weak to expose the photoresist to the degree necessary for it to be washed away during development. But the negative amplitude of this weak phase-shifted light will destructively interfere with the non-phase-shifted light so that a useful (if weak) improvement in the resolution is achieved. The attPSMs are generally considered better than rim-shift PSMs. This

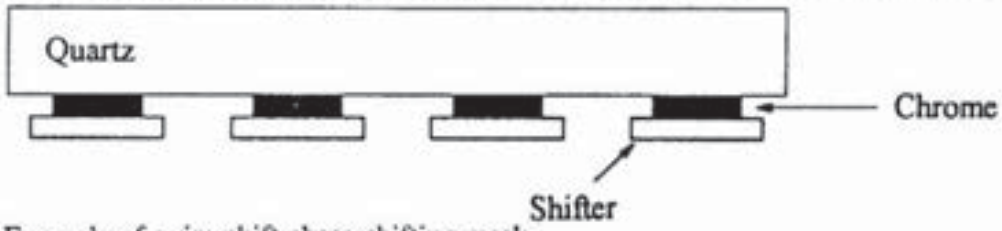


Fig. 7-21 Example of a rim-shift phase-shifting mask.

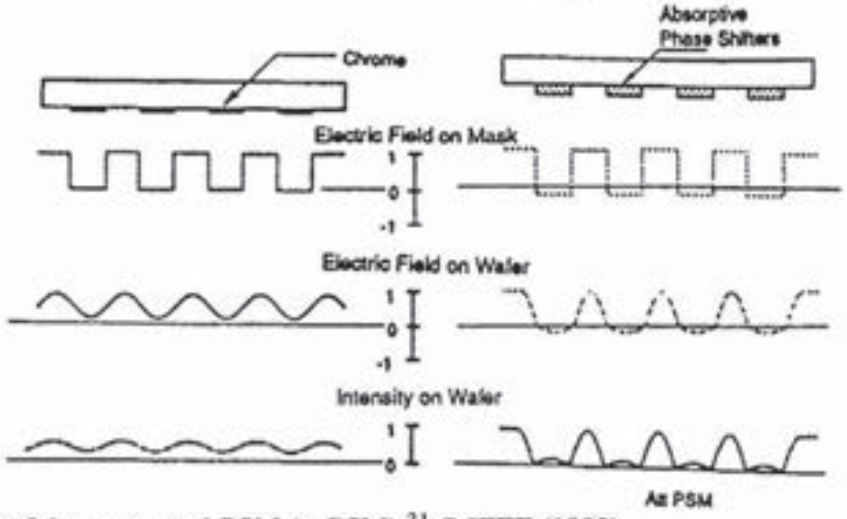


Fig. 7-22 Principle of the attenuated-PSM (attPSM).[31] © IEEE (1993)

is because attPSMs are much easier to fabricate (since the halftone absorber layer can be patterned just as if it was the opaque layer of a conventional mask). Isolated defects (either clear or opaque) can also be repaired with conventional repair techniques. Because of their relative simplicity, attPSMs have been the first type of PSM to be adopted in commercial IC fabrication. One limitation of attPSMs is that they are themselves subject to optical-proximity-effects. One approach to reducing this problem is to combine OPC with attPSMs. The reticle in this case must be made with three tones (clear, attenuated chrome, and opaque chrome). This technique is called *ternary attenuated PSM*, and MicroUnity provides software that allows such OPC features to be designed into the attPSM reticle pattern.[35]

Although PSMs were introduced in 1982, there was scant interest in the technique until about 1988. At that point, a large surge in research about PSMs began. While this resulted in dozens of papers dealing with this topic being published over the next ten years, PSMs were not widely introduced into production until the late 1990s. Even by 1997 only about 1% of the world's mask output contained PSMs. However, as 0.18-μm-CMOS was being moved into production in 1999 it became apparent that resolution-enhancement-techniques including PSMs would have to be used to enable optical lithography to print such small features. By 2000 attPSMs were being used extensively in manufacturing. Such weak attenuating (or halftone) PSMs were being used on 0.18-μm contact-hole levels. Strong-ALT-PSMs were being used on levels containing gates and interconnect lines. ALT-PSMs were also being used in memory circuits, in which interconnect lines are closely and regularly spaced. Reticle inspection techniques for ALT-PSMs were introduced by 1999, but repair techniques still lagged at that

time. The combination of attenuated PSMs and off-axis illumina-tion provides an especially effective resolution-enhancement technology, as each method alone is only able to partially reduce the zero-order light component.

From the descriptions provided above – and from that of previous sections - we can see that a particular type of phase-shifting-mask or modified-illumination is applicable as a RET in specific situations, but not in others. It was noted that ALT-PSMs are effective for dense lines and spaces but not for isolated contacts. On the other hand, attPSMs are applicable for contacts and isolated features. Likewise, off-axis-illumination is found to enhance imaging of dense lines and spaces, but does not particularly improve the patterning of isolated-features. These general trends appear to be true for many of the newly developed methods of enhancing optical lithography. That is, as the improvement is restricted to specific types of features; there is no single solution for them all.

7.4 MASK ERROR FACTOR (MEF, or MEEF)

Masks generally include geometries of various dimensions. The fact that a mask contains a fairly wide range of feature sizes implies that lithographic imaging tools and resist processes should exhibit *linearity*. By this we mean that imaged patterns on the wafer should have the same size as the original patterns on the reticle (after applying the lens reduction factor). In other words, the printed features should all be approximately the size of the corresponding feature on the mask (divided by the lens reduction factor), over a wide range of feature sizes (i.e., a 4X reduction stepper should print a 10-μm-wide line on the wafer from a line on a reticle which is 40-μm wide, and it should print a 1-μm-wide line on the wafer from a 4-μm-wide line on a reticle).

However, as the printed feature sizes approach the resolution limit of the lens, *non-linearity* is found to occur.[37] This effect is shown in Fig. 7-23, which shows the results for a 248-nm stepper. The data in the figure indicates that linearity between the reticle-feature-size and printed-feature-size is obtained for printed-feature-sizes down to about 150% of the exposure-wavelength. That is, one can see that for the case of a 248-nm stepper, as long as the reticle dimensions are larger than 1300-nm (which is 4 x 325 nm), the reticle-dimensions are transferred to the printed-features in a 4:1 ratio. However, for smaller features, the dimensions on the wafer are transferred at a ratio greater than 4:1. This means that when critical geometries are being printed, corrections on the mask must be made to account for this non-linearity. This type of print-bias is a function of feature-size, and therefore can be corrected by appropriately resizing the features on the mask.

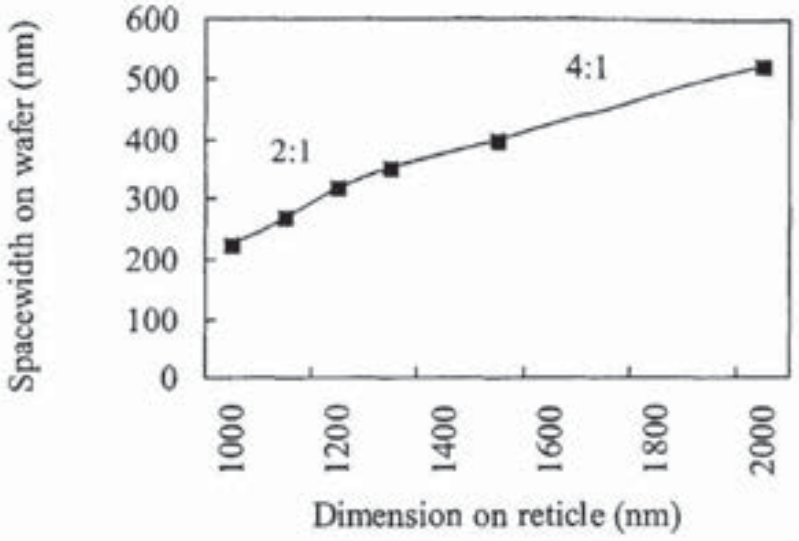


Fig. 7-23 Transfer of features from a reticle to the wafer. The features are reduced in size by 4x for large feature size, but the reduction changes for features near the resolution limit of the lens.

Since the quality of exposure tool lenses has improved and resolution enhancement technology has provided the means to push imaging performance below the wavelength of the exposure light (i.e., to less than 100% of the exposure wavelength), the smallest features that we now expect to be able to print are in the regime where such non-linearity becomes significant. Hence, this non-linearity effect has become very important from a practical point of view.

Another consequence of the non-linear printing of features near the resolution limit of a lens is that there is a *change in the magnitude by which mask linewidth errors* are transferred by the lens. That is, for larger printed feature sizes (i.e., those which exceed about 150% of the exposure wavelength), mask errors are transferred to the printed feature in the same ratio as the lens reduction factor. For example, in a 4X lens, a 20-nm error in the reticle CD for a 400-nm (4-μm) mask feature, would result in only a 5-nm printed error in the 1-μm printed line (because the transfer is still linear for this feature size). However, what is implied by the data of Fig. 7-23 is that *variations in linewidth about their nominal size are not reduced by the lens reduction factor when small features are imaged.* Instead, the effective mask-dimension-error is magnified. The severity of this amplification is described by the so-called *mask error factor (MEF).* Note that this factor is also called the *mask error enhancement factor (MEEF)*, but we will use the abbreviation MEF here. The MEF is defined as:

$$\text{MEF} = \text{M}\,(\Delta\text{CD}_{\text{printed}}/\Delta\text{CD}_{\text{mask}}) \qquad \textbf{(7.2)}$$

Where M is the nominal magnification of the mask compared to the wafer image, and $\Delta CD_{printed}$ is the change in the CD of the printed feature, and ΔCD_{mask} is the change in CD on the mask. If MEF > 1, the errors in the mask CDs are transferred to the printed features on the wafer by an amount that is greater than would occur if MEF = 1, even if the mask CD gets geometrically reduced by the proper ratio (e.g., 4:1). This makes it more difficult to control wafer CDs in the presence of inevitable mask errors. For example, for the case described in the previous paragraph, the MEF = 1. However, if the MEF in that case was instead = 2, the error in the CD of the printed line would increase from 5-nm to 10-nm for an error of 20-nm on the reticle (see Fig. 7-24).

The primary factor that causes the values of MEF to rise above 1.0 is the degradation of the image contrast as the factor ($CD_{printed}$ λ/NA) is decreased below 0.6. In an imaging system with no aberrations and no defocus, it has been found that MEF rises to over 4 when ($CD_{printed}$ λ/NA) = 0.35. (The 0.35 value corresponds to a resolution of 125-nm for a KrF-stepper with an NA of 0.7.) In such circumstances (i.e., if MEF = 4), any reticle linewidth variations will transfer 1:1 on 4X reduction exposure tools. In such cases, the entire benefit of lens reduction (whose role is to reduce the impact of reticle variations on wafer linewidth control) is completely absent.

It has been found that exposure-dose has little significant effect on the MEF. The MEF is related to the mask size and the imaging system, not the resist size.

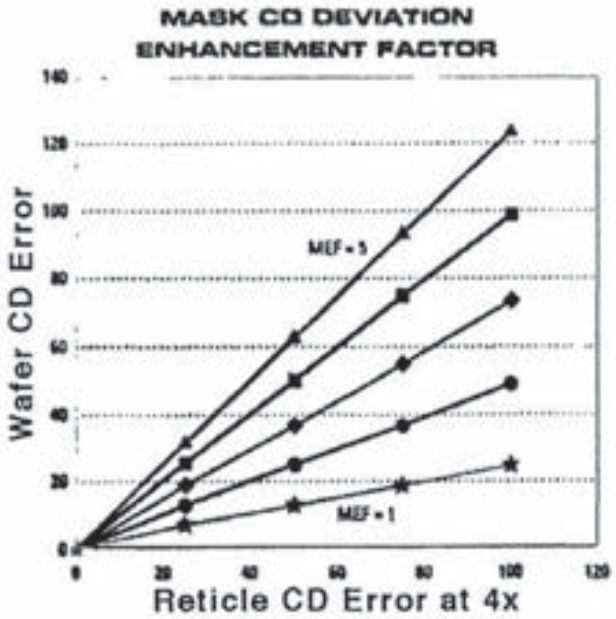


Fig. 7-24 For a MEF >1, a small CD-error on a 4X mask can transfer a larger magnitude of error onto the wafer. (Courtesy KLA-Tencor)

Therefore, reticle biasing does not really help to reduce MEF, although it may be desired for other reasons. The MEF degrades somewhat with defocus, but the degradation is small for defocus smaller than the Rayleigh unit of defocus, $\lambda/2NA$.

The most significant factor that can control the MEF is improvement in the image contrast. This improvement can be achieved in various ways, including the use of more advanced exposure tools (shorter wavelength, higher NA, better aberration control), and through the use of some of the resolution enhancement techniques. Of particular benefit, at small cost, is the use of annular or quadrupole illumination. The disadvantage is that the proximity effects tend to be emphasized and the isolated-to-dense feature bias control suffers.

The MEF also has a significant impact on reticle costs. That is, as long as MEF = 1, improvements in reticle linewidth control only had to match the decrease in minimum feature size (in order to keep constant the reticle contributions to wafer linewidth variations). However, as noted above, with MEF values as large as 4 now being reported, the challenges facing reticle makers have been increased. That is, multiple improvements in the quality of photomasks will be necessary to limit the effects of MEF. Even if these improvements should prove technically feasible, they will only be achieved with increased cost.

These reticle improvements include tighter CD-control and metrology, and tighter enforcement of defect inspection parameters. Characterization must be performed in each area of the wafer because MEF is not uniform (isolated lines generally have a lower MEF than do equally sized dense lines, and near the diffraction limit, contacts have ~2x the MEF of lines and spaces of similar width), and defects near the feature edge are magnified by MEF (just like CD errors). This creates a paradigm shift in sample size and will also require new developments in defect inspection tools. Typically, reticles have been specified with 5 to 10 measurement-sites using white light or *i*-line-metrology. However, with tighter specifications in both CD and defects, the 1-μm optical metrology limit and such small sample sizes are no longer adequate. Generally used at the wafer level, CD SEMs with resolutions ~4 nm, are now being applied to manufacturing the mask sets for 0.15 μm memory cells (whose CDs are 0.6 μm on the reticle). Inspection-tool improvements are allowing such higher resolution limits to be achieved on 5 to 15 times more sites – with overall measurement times remaining the same (~5 min/mask). Another path toward reducing MEF, as described by Kim *et al.*,[38] is to widen the exposure latitude using all means available (including improving aerial image contrast with strong phase-shift masks).

7.5 EXTENDING THE LIMITS OF OPTICAL LITHOGRAPHY

In the late 1970's, the demise of optical lithography was predicted to occur when IC feature sizes reached about 1-μm. Since optical lithography is being used for 0.13-μm IC technologies, this prediction was obviously incorrect. Nevertheless, some limits to optical microlithography do exist, and the consensus today is still that some other lithographic technology will replace optical lithography at a future date.

The resolution limit of optical lithography was also once considered to be about the same dimension as the wavelength of the exposure light. However, as noted in Sect. 7-3, by employing *resolution enhancement technologies* (RETs) the minimum feature size that can now be practically imaged is about 50% smaller than the exposure wavelength (e.g., 0.13-μm features can be produced with 248-nm KrF light).[39] As a result, the rate of shrinkage of the minimum feature size has been faster than the reduction of the exposure wavelength. (For 0.18-μm technology, KrF excimer-laser-based lithography is employed, with the KrF wavelength being 248-nm [0.248μm]).

To image 0.13-μm features without having to resort to such resolution enhancement techniques as PSM, off-axis-illumination, OPC, and multi-layer resists, however, a shorter wavelength of light is needed (e.g., ArF-excimer-laser light that has a wavelength of 193-nm). Since ArF technology (including appropriate light sources, lens materials, and resists), did not become available in time for full-scale production of 0.18-μm ICs, adoption of such RETs onto KrF has been the way that 0.18-μm technology is being manufactured. In 1999 the first generation 193-nm ArF 4:1 stepper/scanners from ASML, Canon, and SVG Lithography were installed on 0.13-μm development lines. The average price of ArF steppers is \$10-\$15 million. ArF resists from Olin, and others (based on acrylates and cyclic olefin polymer materials – see Chap. 6), were available for pilot-line use. But it appears that ArF excimer-light (or some other shorter wavelength energy source) will be needed to obtain adequate resolution with sufficient depth of focus for the next generation (0.1-μm-CMOS). Note that even at 0.1-μm, the dimension of the feature being printed is smaller than even the wavelength of ArF excimer-light. Thus, RETs will also be needed. (Another financial drawback is that ArF light appears to damage the fused-silica used for ArF lenses. Thus, the \$1.5-million lens in ArF stepper/scanners may need to be replaced after three years of use.) Since (for reasons we will describe next), it appears that the limit of optical lithography is now predicted to be 0.05-0.07-μm (50-70-nm), non-optical techniques will eventually be needed.[45] Such non-optical alternatives are considered in the next section.

We noted above that in the 1970's it was predicted that it would not be possible to use optical lithography for IC technologies below about 1 μm. There were several reasons why this estimate proved to be incorrect, as follows:

1. In the decade starting in 1985, lens-making technology made significant and unexpected advances. For example, as noted in Sect. 7.2, 0.3-NA, *g*-line lenses being manufactured in 1980 had a resolution limit of 1.25-μm and a depth-of-focus of 1.5-μm. By 1996 Nikon was offering lenses with the same NA, but with much improved performance (i.e., these Nikon *i*-line-lenses with 0.3-NA had a resolution limit of 0.7-μm and a depth of focus of 5.0-μm). The reason for this improvement was that the lenses being manufactured in the mid-1985's were not yet diffraction limited (but were still being limited by lens aberrations). By 1996, lens-making technology had improved to the degree that the lenses actually were diffraction limited. In addition, it was learned how to make lenses with much larger NA than in 1980. For example, by 2001, lenses with an NA as large as 0.8 were being offered commercially. (However, since the NA of lenses cannot exceed 1, and since diffraction-limited lenses are as perfect as they can be, it will not be possible to obtain the gains provided to optical lithography from such ingenuity and innovation in lens-making again.)

2. Optical lithography technologies were developed that could use shorter wavelengths than the mercury *g*-line (436-nm). The ability to use such shorter wavelengths not only reduced the resolution limit, but also brought with it an increased depth of focus.

3. Photoresists also were improved, allowing the k_1 values to be reduced. For example, in 1980 $k_1 = 0.8$, but by 2001, k_1 was decreased to < 0.5.

4. The depth of focus requirements (DOF) were decreased (down to about ± 0.2-μm), primarily due to the development of chemical-mechanical polishing (CMP).

5. Resolution enhancement technologies (RETs) were developed.

Despite these advances, there are several physical limits that are being reached that imply that the optical lithography will need to be replaced when the minimum feature sizes on ICs reach 50-70-nm. Let us list those limits, which have been stated by such industry experts as H. Levinson to justify the 50-70-nm resolution limit of optical lithography.[40]

1. To get to 70-nm resolution with optical lithography, shorter wavelengths of light will be needed. UV lasers operating at 157-nm (F_2) and 126-nm (Ar_2^*), are possible illumination sources for such shorter wavelengths.

Since virtually all materials absorb strongly in the range of 20–193-nm (except for some, like calcium fluoride [CaF_2] that still transmit 157-nm and 126-nm light), it may be necessary to use all-reflective optical systems at these wavelengths. (Extreme UV [EUV] lithography, discussed in Sect.7.6.1, uses such all-reflective optical systems.) However, as discussed previously, by 1999 advances in F_2-laser technology, in the growing of CaF_2-crystals, and in 157-nm resist chemistry made 157-nm lithography a possible candidate to fill the 0.1-μm "lithography gap" that exists.[41]

However, the thermal stability of the mask-substrate material may make it impractical to go to 126-nm illumination. That is, fused-silica can be still used for the mask substrate material because it still sufficiently transparent at 157-nm. However, it becomes too absorbing just a few nanometers below 157-nm, and thus it is not a material that can be used for masks at 126-nm. For such masks, a material such as CaF_2 must be used (since CaF_2 is still sufficiently transparent at 126-nm). However, the coefficient of thermal expansion of fused-silica is much smaller than that of CaF_2 (0.5-ppm/K vs 19 ppm/K). When a reticle is being exposed with an e-beam (to create the patterns on the reticle), substantial energy is absorbed by the mask-substrate material, and this causes the temperature of the mask blank to rise. Assuming that the temperature-rise can be limited to only 0.1K, there will be a 5-nm registration error across 100-mm on a substrate made

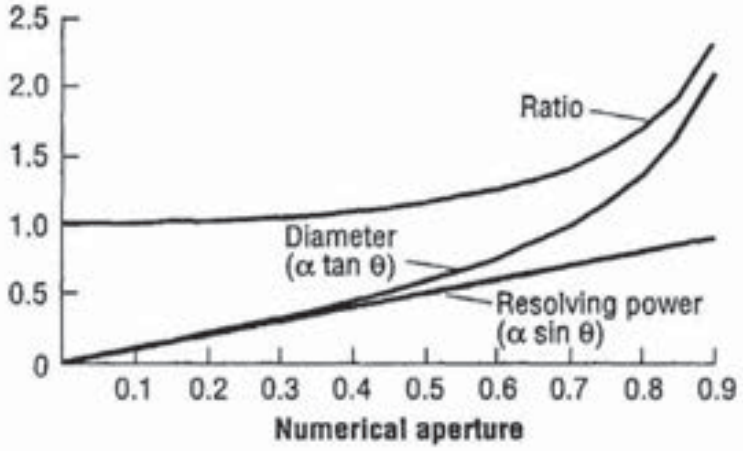


Fig. 7-25 Diminishing returns of higher NA for higher resolution. For lens with an NA greater than 0.80 NA, each incremental increase in lens-resolving power requires disproportionately larger increases in lens diameter. The ratio represents the difference of lens diameter to resolving power at a given NA.

of fused-silica (a magnitude of mis-registration that can easily be accommodated within overlay budgets). However, if the mask blank is made of CaF_2, the same temperature-rise will cause a mis-registration error of 100-nm across 100-mm, making it impractical to use CaF_2 as the mask substrate material. Thus, it appears that 157-nm will be the shortest wavelength that can be used for an optical lithography process.

2. As noted earlier, lenses cannot be made any more perfect than when they are diffraction limited. Thus, the remaining limitation imposed by diffraction remains and becomes dominant. In addition, it is unlikely that lithography will be carried out in a medium other than a vacuum or gas (such as air or nitrogen). This limits the NA of a lens to < 1. Since the NA of lenses available in 2001 is already at 0.8 in the most advanced scanners, increases in NA in the future can no longer be a key driver to higher resolution that it has been in the past. Thus, the great strides of lithography due to improvements in lens fabrication are no longer possible (Fig. 7-25).[39]

3. The decrease in depth-of-focus requirements down to ±0.2-μm was brought about by the use of CMP. However, it will not be possible to reduce this requirement further, as the wafer topographies are now about as flat as they can get.

4. The degree to which optical lithography can be extended toward the 50-nm limit depends on how effectively all of the RET methods described earlier can be implemented. The ability to implement a particular RET technique also depends on the application. As an example that was described earlier, the use of off-axis-illumination is highly effective for imaging highly repetitive structures (such as those found in memories), but the same technique is of not much benefit for imaging isolated features.

In summary, there are many challenges to be met to enable optical lithography to be extended to 70-nm (0.07-μm), although there is optimism that this goal will be achieved (Fig. 7-26).[40] However, for the reasons given above, 50-nm (0.05-μm) appears to be a clear lower limit for optical lithography.[41]

It is also probable that a mix-and-match strategy will be employed when ICs are fabricated with features smaller than 0.1-μm. That is, a non-optical lithography process may be used for the layers with the smallest minimum dimension, and optical lithography techniques for the non-critical layers. Figure 7-27 shows the relationship between resolution and exposure wavelength for various types of microlithographic technologies.[42]

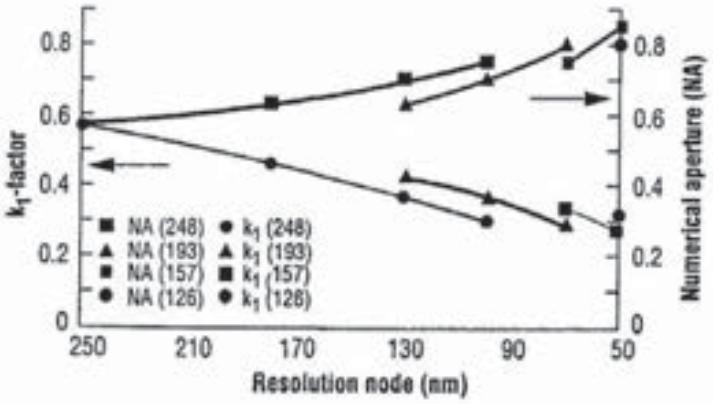


Fig. 7-26 Wavelength, NA, and k_1 for 250 nm to 50 nm resolution nodes. (Courtesy ASML)

Overlay limits are another issue to be considered with optical systems. Stage precision is the biggest practical concern, but work is reportedly being carried out to solve this problem.[43] Another overlay issue involves alignment target acquisition. Metal and grainy polysilicon have long presented difficulties for stepper

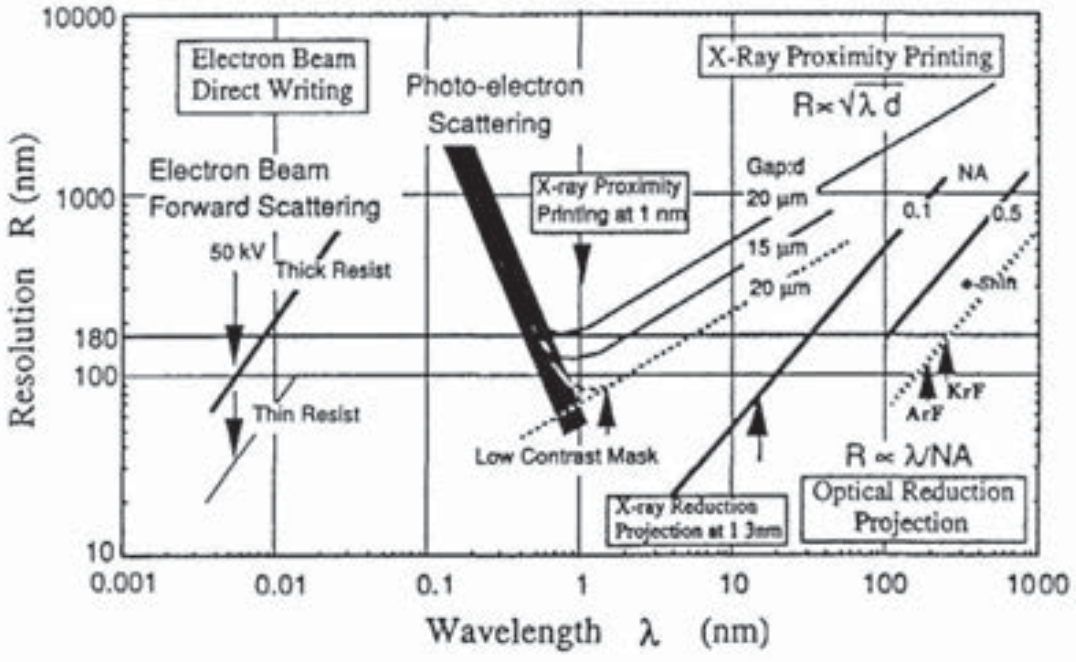


Fig. 7-27 Relationship between resolution and exposure wave length for various microlithographic technologies.[42] © 1996 (IEEE).

alignment systems. More recently, additional problems have been encountered on wafers that have undergone chemical-mechanical polishing. One solution has been to align all layers to a "zero layer" having consistent alignment targets.

The best level-to-level overlay-capability of optical steppers is estimated to be about 30-nm, which will be adequate for 0.1-μm technology, but not for 0.07-μm. For 0.07-μm technology, it may be necessary to dedicate individual steppers to a particular wafer lot. This will reduce overall stepper productivity and introduce another source of additional costs.[44]

7.6 NON-OPTICAL (or NEXT GENERATION) LITHOGRAPHIC TECHNOLOGIES (NGL)

In order to image features smaller than 70-nm it will probably be necessary to employ some form of non-optical technology (or *next generation lithography, NGL*).[52] Two such NGL candidates will be discussed in this section: **1**) extreme ultra-violet (EUV) lithography; and **2**) electron-beam projection lithography (EPL). It should be noted, however, that by 2001, extreme ultra-violet (EUV) lithography had become the leading NGL technology. The others had become extinct (proximity x-ray and direct-write e-beam), or on put on the endangered species list (EPL). The EPL approach suffered a punishing blow in late-2000 when Agere Bell Laboratories cancelled it's research effort. However, since it remains alive under the auspices of an IBM-Nikon led research effort, we will provide a brief description of this candidate NGL technology as well.

It should also be noted that in 2001 the consensus was that non-optical techniques were still behind the overall production capability of optical micro-lithography (especially with respect to productivity).[45] Two additional difficulties are shared by non-optical lithographic techniques: **1**) the cost of a single mask is expected to be in the $50,00–$100,000 range; and **2**) no pellicles can be mounted on the masks to protect them from particles. The latter problem makes masks for non-optical technologies much more vulnerable to particles and other defects.

7.6.1 Extreme Ultra-Violet Reflective Projection Lithography (EUV)

An attempt is being made to extend the principles of projection lithography into the soft x-ray spectrum. (The region of the spectrum with photon energies of about 100-eV and with wavelengths between 10-15-nm [or 0.01–0.015-μm] can be considered either at the low end of the x-ray spectrum (*soft x-rays*), or the short wavelength limit of the *extreme ultra-violet* spectrum [EUV].) The chief reason EUV seems to be emerging as the winner in the NGL race is that EUV demands the fewest changes in manufacturing from the current methods, and

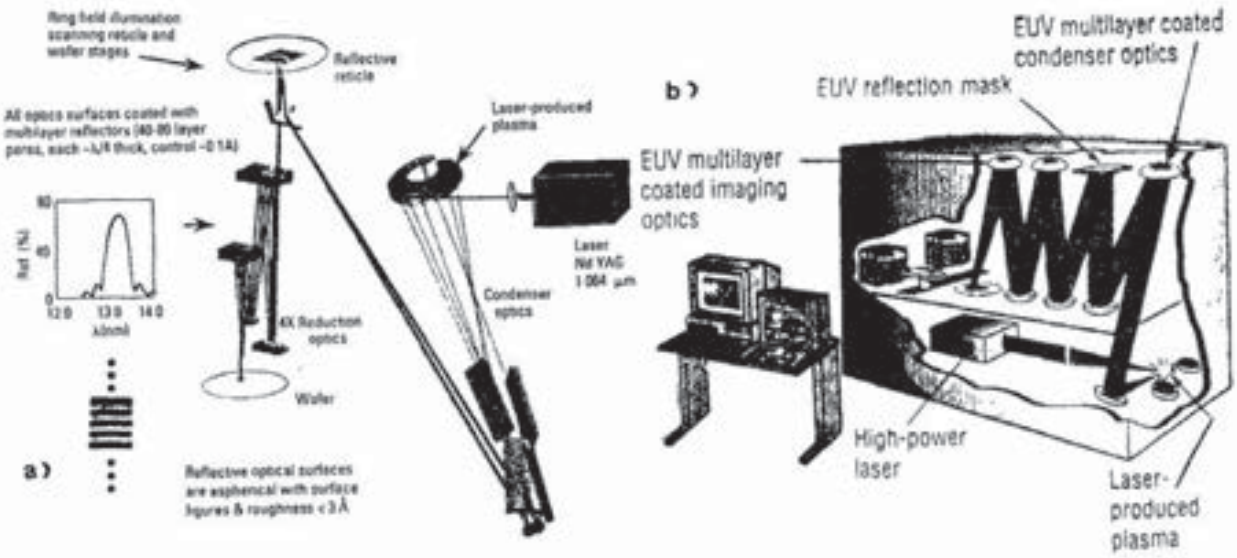


Fig. 7-28 (a) Schematic drawing of the optics of an EUV system. (b) An artists drawing of an EUV lithography system. A laser generates EUV radiation by exciting a laser produced plasma. This radiation is collected by a multilayer condenser system. The radiation illuminates a multilayer coated reflection mask, which is imaged and demagnified onto a resist-coated wafer.[46] Reprinted with permission of Solid State Technology, published by PennWell.

therefore appears to involve the least risk. (In an industry where downtime at an average fab costs $4 million a day, risk is not taken lightly.) For example, EUV radiation is more like light than any of the other alternatives. That is, it still involves photons, rather than electrons or ions, and the industry has grown comfortable with photons.

Furthermore, EUV photons can be focused with reflective lenses (which are actually curved mirrors). Use of these mirrors enables reduction optics to be used, which in turn allows photomasks to be fabricated with much larger dimensions than the features being imaged. Such masks are much easier to fabricate than masks containing patterns of the same size as the imaged features.

A number of interference coatings that exhibit relatively good reflectivity of such EUV light have been developed (i.e., silicon/molybdenum composites - deposited by sputtering), and they can be used to produce EUV mirrors. These coatings make it feasible to design projection systems that use all-reflective lithographic components to image patterns with EUV. The microlithographic technology based on this approach is called *extreme ultraviolet lithography*

(*EUVL*), or *soft x-ray projection lithography*.[47] Figure 7-28 shows a schematic of an EUV lithography system. A short survey of EUV technology is provided here. For more details, readers are referred to the excellent book by Attwood.[53]

EUV radiation can be generated in a number of ways. *Laser plasma sources* have shown promise as EUV light sources. They utilize pulses of high-intensity laser-light (normally from a Nd-YAG laser) to strike a target material (typically a supersonic jet of xenon gas). This produces a plasma that emits light with a broadband radiation that contains a significant component of EUV radiation. Xenon gas is the most common target material because it produces relatively strong EUV radiation, and being inert, does not directly generate debris that would contaminate and degrade the EUV optics. Unfortunately, these lasers may not be robust enough to be able to provide reliable operation in a production environment.

Synchrotrons (such as those used in proximity x-ray lithography) can also serve as the source of EUV light. Their main drawback is that they are not standalone light sources. Chip makers would prefer a standalone light source. With many standalone light sources in a fab instead of one synchrotron, the risk of having to shut down an entire fab if the light source fails is avoided. Some other alternative light source technologies were reportedly under development, but by the end of 2001 public information about them was not yet available.

In any case, the EUV radiation from the light source is collected and shaped into a narrow arc by a system of mirrors coated with EUV interference films. This radiation is then used to illuminate an EUV reflection reticle. The pattern on the reticle is then de-magnified (by 4X or 5X) and imaged onto a resist-coated wafer. The entire reticle pattern is exposed onto the wafer by synchronously scanning the reticle and the wafer (i.e., a step-and-scan exposure is performed). By using Rayleigh's resolution equation, it is predicted that a resolution of 32-nm can be achieved using a wavelength of 13.4-nm in a system having an NA = 0.25, with a k_1-value of 0.6. EUV lithography also offers the benefit of a large depth-of-focus for a system of such resolution (i.e., a DOF of greater than 1-micron). As noted earlier, since reduction optics are possible, many of the mask problems associated with proximity x-ray lithography are overcome.

However, there are a number of formidable challenges that must be met to make this technology practical. First, a modestly-priced, highly-reliable light source is still needed. Second, the stepper and the EUV beam-delivery system must be operated in a vacuum to prevent absorption and scattering of EUV rays. Third, the absence of transparent materials at EUV wavelengths implies that pellicles cannot be used (especially since the EUV light must pass through a

pellicle twice in the reflection optics of EUV systems). Finally, since there are no high-transmission materials in the EUV, not only must all optical elements be reflecting, so must the mask. There are numerous issues involved in the fabrication of such reflecting optical elements and masks, and these will be discussed in the following paragraphs. While all of the above challenges appear daunting, significant progress was reportedly being made by 2001 to meet them.[55,56]

Fabrication of the reflective optical elements and masks for EUV requires a technology that can not only create surfaces with a relatively high reflectance, but that can also shape the surfaces of such reflective optical elements to extremely high optical tolerances. High mirror reflectance is important for EUV exposure tool productivity because such tools contain many reflecting surfaces (e.g., six mirrors in the projecting optics, two mirrors in the illuminator, and a reflection mask). Since the EUV light reflects nine times in such a system, the intensity of light at the wafer plane will be nearly 30% greater if each reflector has a 70% reflectance - compared to the case where each reflector has a 68% reflectance. Hence, even small increases in reflectance are significant.

Reflection occurs at interfaces between materials of different indices of refraction. However, at wavelengths below 50-nm all materials exhibit comparable indices of refraction (~1). Thus, it is difficult to make *single* interfaces that are good reflectors of EUV light. However, a technique that uses *multiple* interfaces can produce moderately effective EUV reflectors (i.e., with 60-70% reflectance). That is, such reflectors consist of a multilayer structure having a stack of many (e.g., 40 pairs) thin films (i.e., each with a thickness of tens of angstroms). This stack consists of alternating thin layers of high-Z and low-Z materials (where Z is the atomic number of the element being used in the film). This arrangement is used because there is a small but effective difference between the refractive indices at a high-Z/low-Z interface (see Fig. 7-29). Furthermore, by making the periodicity d of the multilayer satisfy the Bragg condition ($\lambda/2\sin\theta$), the reflected light from each interface adds to the reflected light from the other interfaces. The net effect is that by combining the small reflectivity at each interface it is possible to achieve a moderately high reflectivity overall (assuming there are a sufficient number of layers).

The materials for each layer must also be weak absorbers of EUV light, since the light must be able to penetrate to the lower levels of the film stack (which also explains the need to use as thin a film as possible for each layer). Most efforts have focused on using Mo and Si as the high-Z and low-Z materials for each layer of the EUV reflector, respectively. A 68% reflectance has been

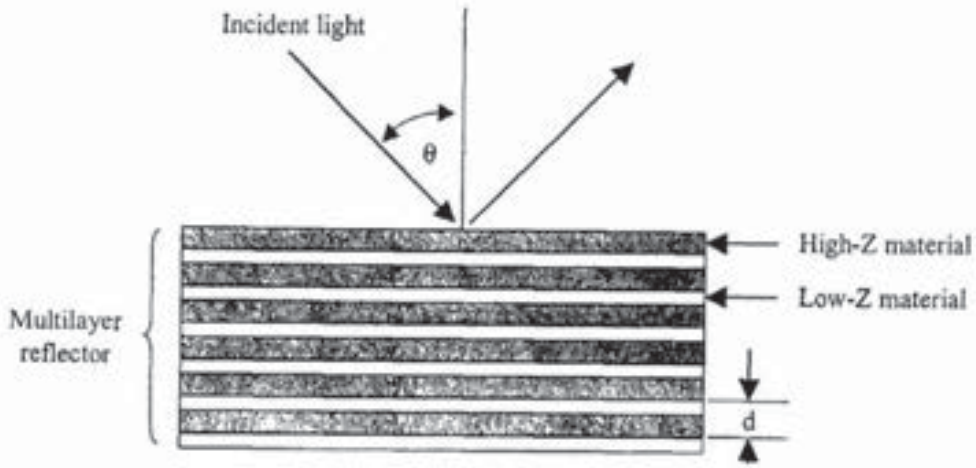


Fig. 7-29 Principle of a multilayer reflector for EUV radiation.

achieved for reflectors made with such Mo/Si multilayers for an EUV wavelength of 13.4-nm. However, the reflectivity will need to be improved to >70% to enable high-throughput cost-effective lithography. These layers are deposited by physical vapor deposition (sputtering), and each layer must be deposited to a thickness accuracy of less than an angstrom. Note that since most elements are strong absorbers at EUV wavelengths, even sub-monolayers of contaminants can reduce mirror reflectance. As a result, mirror (and mask) contamination is an important issue for EUV lithography. Finally, these interference films must be able to withstand exposure to strong EUV radiation without damage over long periods of time.

The fabrication of the aspheric (non-spherical) mirror substrates (the curved surfaces onto which the multilayer stacks must be deposited) is difficult because of the extremely tight tolerance (i.e., better than 2-nm) on the *surface figure* of these elements. Optics producers will need new metrology techniques to measure figure and surface roughness of these mirrors to the required accuracies. It was reported by workers at the Lawrence Livermore Laboratories in March 2001 that it was taking a full year to make a mirror set for EUV at that time. Industrial insertion would require manufacturing 300 such sets/year, a daunting challenge.

The *EUV masks* are fabricated by depositing the reflecting multilayers described above, on a flat substrate. Absorbing materials are then deposited and patterned on the multilayer, completing the mask fabrication. Since nearly all materials absorb EUV radiation, various films can be used for the absorber,

including chromium, TiN, and TaN. The choice of absorber material is based on their ability to be deposited, etched, and cleaned. One of the problems of using Mo/Si multilayers on EUV masks is that they begin to significantly intermix at temperatures above 150°C. Such intermixing reduces the reflectivity of the stack. This limits the maximum process temperatures used to deposit and etch the absorber layers on EUV masks. By 2001, it was reported that EUV developers had devised a way to repair masks when defects are inadvertently incorporated during their fabrication.

Another consequence of the high absorptivity of all materials at EUV is the need for ultrathin resists (< 200-nm-thick). The resist-collapse problem that occurs when resist features have high aspect-ratios also dictates the use of thin resists (Chap. 6). However, it has been demonstrated that lithography processes suitable for manufacturing are feasible with such ultrathin resists. Moreover, resists quite similar to those used for DUV lithography appear to be suitable for EUV lithography.

EUV development is being carried out by a unique U.S. government-industry consortium known as the EUV Limited Liability Corporation (LLC). Under the terms of the agreement, three national laboratories, Sandia, Lawrence-Livermore, and Lawrence-Berkeley – are doing most of the development work, while industry members supervise and foot the bills. The core of the industry group consists of five chip makers: Intel, AMD, Infineon, Micron Technology, and Motorola. In April, 2001, the first full-scale prototype EUV lithography system was unveiled (Fig. 7-30).[54] This system, called the Engineering Test Stand (ETS) measures about 10x10x10 ft. The light source is a 1500-watt laser plasma source that generates an EUV beam with 10-W of power. (More than 50-W will be needed by the time the EUV systems reach production, as this is the power expected to be needed to allow throughputs of 70 300-mm wafers per hour.) The EUV LLC announced in March 2001 that the ETS was used to successfully print lines at 100-nm spacings. It was predicted that by the end of 2001, the ETS would be printing 70-nm lines and spaces at full field sizes. Debates are occurring at the end of 2001 as to when EUV will be ready for high-volume production. The optimists say that this will take place by 2005. Others argue that while early production units may be ready by then, high-volume manufacturing will take 2 to 3 years longer.

7.6.2 Electron Beam Projection Lithography (SCALPEL and PREVAIL)

Another NGL technique based on *e-beam projection* has been pursued (EPL). The system developed by Lucent Technologies (now Agere), is called *Scattering*

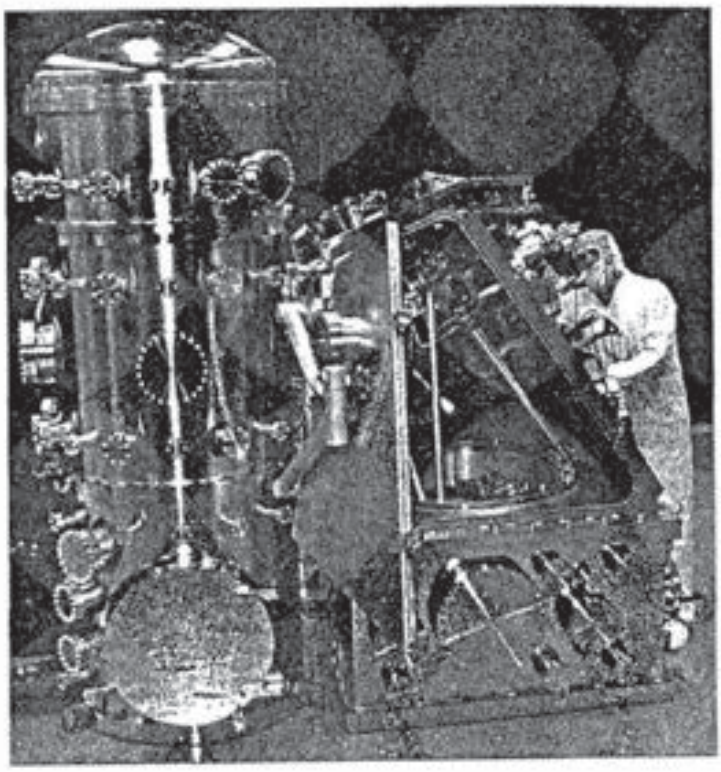


Fig. 7-30 A worker inspects the Engineering Test Stand, a full-scale prototype EUV lithography system – as assembled at Sandia National Laboratories. (Courtesy Sandia Labs).

with Angular Limitation in Projection Electron beam Lithography (or *SCALPEL*).[47,48] In this approach, a special scattering mask is used to produce a high-contrast image with a technique commonly used in transmission electron microscopy. That is, an electron beam (coming down from the top of Fig. 7-31) is scanned across a special mask that consists of a thin silicon nitride (SiN) blank (100–150-nm thick) having mask patterns formed of very thin (i.e., 25–50-nm-thick) tungsten or gold features. The electrons from the scanned beam can either: **1**) strike regions of the mask covered with one of the metal patterns; or **2**) strike a region of the mask *not* covered with metal. Electrons striking the SiN film pass straight through it (i.e., SiN only weakly scatters the beam electrons). On the other hand, electrons striking the metal patterns are strongly scattered as they pass through the heavy metal of the mask features. Those that have passed unscattered through the silicon nitride are then focussed by a lens so that they pass through the aperture of the *back focal-plane filter*. But most of the electrons that are scattered as they travel through the metal patterns get blocked by the back focal-plane filter (and thus do not travel though its aperture). The electrons that make it through the filter aperture are demagnified and strike the wafer

surface (whereupon they form a high-contrast, 4:1-reduced image of the scattering mask).

The resolution limits of SCALPEL can be extended below 0.1-μm. In addition, because the scattering patterns are so thin, the demagnification of the mask can be made small enough that chips can be fabricated with as many as $2x10^{10}$ pixels (defined here as a small square whose side is equal to the minimum printable linewidth). Since 1-Gigabit DRAM patterns, for example, need roughly 10^{10} lithographic pixels per chip, the SCALPEL technique is suitable for such chip designs. The thin silicon nitride mask-blank substrate is supported on a silicon wafer with periodic silicon support struts. These struts are not imaged onto the wafer since the patterns are shifted into place as they are illuminated. While the mask structure is similar to that used in x-ray masks, the silicon support struts provide greater dimensional stability of the patterns on the

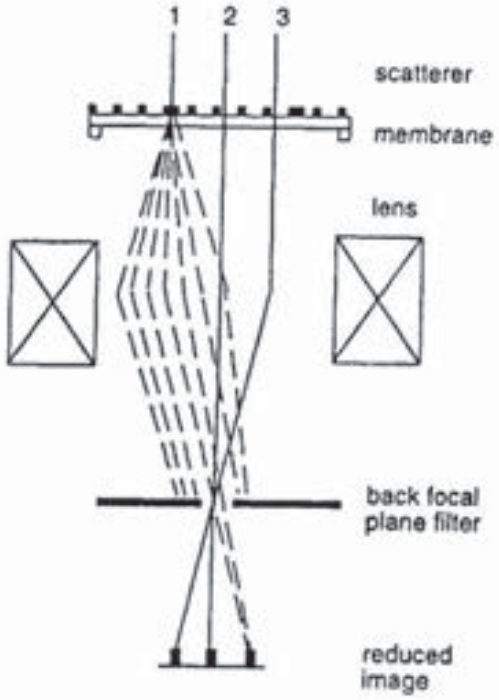


Fig. 7-31 Schematic of the SCALPEL technique. Electrons (1) that hit the scatterer (the metal patterns on the mask) are strongly scattered, and most are filtered out by the aperture. Electrons traveling through the silicon nitride membrane (2, 3) are only weakly scattered and pass through the aperture. These are demagnified to form a high contrast image on the substrate. Courtesy of Lucent Technologies, Inc.

SCALPEL mask, and the use of 4:1 reduction optics makes mask fabrication of the 4:1 reduction masks simpler. The initial price of a SCALPEL system (1998) was $8 million.

One advantage of SCALPEL over *x*-ray lithography is that it does not require a synchrotron orbital radiation (SOR) ring. Such an SOR ring is expensive, and is designed to be used with multiple steppers. Instead, a good e-beam source for SCALPEL is available. For SCALPEL, if another stepper is needed, only the stepper need be purchased. Another advantage is that the scanning e-beam that passes through the mask produces little heat in the thin nitride mask membrane and this reduces distortion of the imaged pattern on the wafer. Finally, SCALPEL's depth-of-focus has been measured to be ~50 times greater than for 0.25 μm optical lithography techniques. On the other hand, the traditional problems of e-beam lithography, such as low throughput, proximity effects, radiation damage, charging, and lack of a good resist are limitations of SCALPEL. Some other concerns involve keeping the membrane clean, and finding ways to repair it. A goal of producing working SCALPEL steppers that can deliver up to 50 300-mm wafer per hour at 0.13-μm resolution by 2003 has been established.

In 2001, however, EPL appeared to be losing some of it's momentum, with EUV gaining support instead. Motorola dropped out of the group of companies backing EPL, and Applied Materials and ASML disbanded their eLith LLC venture that was to have commercialized the SCALPEL technology. However, several of the EPL backers remained committed to developing the approach. Agere and TI switched their allegiance to the EPL approach being developed by Nikon and IBM (as did the semiconductor companies belonging to the Selete equipment consortium in Japan). The Nikon-IBM approach is called PREVAIL. These companies argue that EPL is better-suited to small-volume ASICs than the EUV system, which requires more expensive masks. Nikon is preparing to offer the PREVAIL e-beam column for e-beam projection lithography by 2004 (and perhaps earlier).

REFERENCES

1. O. Semprez, "Excimer Lasers for Future Lithography," *Solid State Technology*, July 2000, p. 255.

2. A. Ershov *et al.*, "Feasibility Studies of Operating KrF Lasers at Ultra-Narrow Spectral Bandwidths for 0.18 μm Linewidths," *Proc. SPIE.*, Vol. 3334, Feb. 1998.

3. A. Hand, "KrF Lithography Not Finished Yet," *Semiconductor International,* February 2001 p. 58.
4. M.D. Levenson, "Is There Light at the End of NGL Tunnel?" *Solid State Technology,* May 2001, p. 51.
5. R. DeJule, "157-nm Lithography Moves Forward," *Semiconductor International,* July 2000, p. 82.
6. R. DeJule, "NGL's Critical Bridge: 157 nm," *Semiconductor International,* March 1999, p. 36.
7. A. Hand, "Ready or Not, When Will 157-nm Lithography Really be Needed?" *Semiconductor International,* May 2001, p. 42.
8. B. Wang, "Birefringence in Fused Silica and CaF_2 for Lithography," *Solid State Technology,* February 2000, p. 77.
9. A. Hand, "Intrinsic Birefringence of CaF_2 Won't Halt 157-nm Lithography," *Semiconductor International,* September 2001, p. 42.
10. "Lithography Tools for 300-mm," *Solid State Technology,* September 2000, p. 40.
11. P.M. Ware, "300-mm Lithography Enters Reality," *Semiconductor International,* July 2000, p. 128.
12. A. K.-K. Wong, *Resolution Enhancement Techniques in Optical Lithography,* SPIE Press, Bellingham, WA, 2001.
13. R..DeJule, "Update on Resolution Enhancement Techniques," *Semiconductor International,* August 2000, p. 54.
14. C.A. Mack, "Understanding Focus Effects in Submicron Optical Lithography: A Review," *Optical Engineering*, Vol. 32, No. 10 (October 1993, pp. 2350-2362).
15. C. Mack, "Optimum Stepper Performance By Image Manipulation" *KTI Microelect. Sem.* p. 209 (1989).
16. D.L. Fehrs *et al.*, "Illuminator Modification of an Optical Aligner" *KTI Microelect. Sem.* p. 217 (1989).
17. M. Noguchi *et al.*, "Subhalf Micron Lithography System with Phase-Shifting Effect," *Proc. SPIE*, Vol. 1674, p. 92, (1992).
18. W. Partlow *et al.*, "Depth-of-Focus and Resolution Enhancement for *i*-line and Deep-UV Lithography Using Annular illumination," *Proc. SPIE*, Vol. 1927, p. 137 (1993).
19. S. Asai *et al., J. Vac. Sci. Technol.* **B 10(6)**, 3023 (1992).
20. K. Toh *et al,* "Three-Dimensional Simulation of Optical Lithography," *SPIE,* Vol. 1463, p. 402 (1991).
21. K. Tounai *et al..,* "Optimization of Modified Illumination for 0.25 μm Resist Patterning," *Proc. SPIE,* Vol. 2197, p. 31 (1994).
22. T. Ogawa *et al.,* "Effective Light Source Optimization with the Modified Beam for Depth-of-Focus Enhancements," *Proc. SPIE,* Vol. 2197, p. 19 (1994).

23. H.J. Levinson and W.H. Arnold, "Optical Lithography," Chap. 1, pp. 11-138, in *Handbook of Microlithography, Micromachining, and Microfabrication,* Vol. 1, Microlithography, Ed. P.Rai-Choudhury, SPIE Optical Engineering Press, Bellingham, WA., 1997.
24. R. DeJule, "OPC Simulation," *Semiconductor International,* September 1997, p. 56.
25. C. Mack, "Evaluating Proximity effects Using 3-D Optical Lithography Simulation," *SemiconductorInternational*, July 1996, p. 237.
26. C.A. Mack and P.M. Kaufman, "Mask Bias in Submicron Optical Lithography," *J. Vac. Sci. Technol.,* Vol. **B6** (Nov./Dec. 1998) p. 2213.
27. N. Shamma *et al.,* "A Method of Correction of Proximity Effect in Optical Lithography," *KTI Microlithography Seminar Interface 1991*, p. 145.
28. *Microlithography: Science and Technology,* J.R. Sheats & B.W. Smith, Eds., Marcel Dekker, New York, 1998.
29. M. Levenson *et al.,* "Improving Resolution in Photolithography with a Phase-Shifting Mask," *IEEE Trans. Electron Dev.*, **ED-29**, 1828 (1982).
30. B.J. Lin, "Phase-Shifting Masks Gain an Edge," *IEEE Circuits & Devices Mag.*, March 1993, p. 28.
31. *ibid.*, Ref. 24, p. 74-82.
32. M.D. Levenson, "Strong Phase-Shift Mask Designs with Lower Cost," *Solid State Technology,* January 2001, p. 104.
33. T. Terasawa *et al.,* "Imaging Characteristics of Multi-Phase-Shifting Mask," *SPIE,* Vol. 108, 25 (1989).
34. B.J. Lin, "The Attenuated Phase-Shift Mask," *Solid State Technology*, January 1992, p. 43.
35. R. DeJule, "OPC for Attenuated Phase-Shift Masks," *Semiconductor International,* May 1998, p. 44.
36. M.S. Hibbs, "System Overview of Optical Steppers and Scanners," in *Microlithography: Science and Technology*, Eds. J.R. Sheats and B.W. Smith, Marcel Dekker, New York, 1998, Ch. 1, pp. 1-107.
37. W. Maurer *et al.,* "Pattern Transfer at k_1 = 0.5: Get 0.25 μm Lithography Ready for Manufacturing," *Proceeding of SPIE*, Vol. 2726, p. 113, 1996.
38. B.-G. Kim *et al.,* "Beneath the MEEF," *Solid State Technology,* August 2000, p. 107.
39. S. Okazaki, *Proc. of SPIE*, Vol. 2440, p. 18 (1995).
40. H.J. Levinson, *Principles of Lithography*, SPIE Press, Bellingham WA, 2001.
41. J.A. McClay and A.S.L. McIntyre, "157 nm Optical Lithography: The Accomplishments and the Challenges," *Solid State Technology,* June 1999, p. 57.
42. S. Okasaki, "Lithography Prospects for 0.18 μm Technology & Beyond," *Tech. Dig. IEDM,* 1996, p. 57.

43. S.A. Lis, "An Air-Turbulence Compensated Interferometer," *Proc. SPIE*, Vol. 2440, 9891 (1995).
44. *ibid.*, Ref. 23, p. 124.
65. T. Brunner, "Pushing the Limits of Lithography for IC Production," *Tech. Dig. IEDM*, 1997, p. 9.
66. A.M. Hawryluk *et al.*, "EUV Lithography," *Solid State Technology*, July 1997, p. 151.
67. S. Berger *et al.*, SPIE Proceedings, Vol. 2322, p. 434 (1994).
68. L.R. Harriott *et al., J. Vac. Sci. Technol.*, **B14(6)**, 3825 (1996).
69. P.M. Ware, "Why Optics is Forever in Lithography," *Solid State Technology*, May 2000, p. 192.
70. C. Wagner *et al.*, "The Technical Considerations of Extending Optical Lithography," *Solid State Technology*, September 2000, p. 97.
71. J.S. Petersen, "Integrated Imaging a Key Approach to Optical Lithography at 100-nm+," *Solid State Technology*, January 2001, p. 80.
72. R. DeJule, "Next Generation Lithography Tools," *Semiconductor International* March 1999, p. 48.
73. D.T. Attwood, *Soft X-rays and Extreme Ultraviolet Radiation, Principles and Applications*, Oxford University Press, 2000.
74. Hand, "EUV Lithography Makes Serious Progress," *Semiconductor International*, June 2001, p. 54.
75. J. Canning, "Challenges of Next-Generation Lithography," *Semiconductor International*, September 2000, p. 107.
76. M.D. Levenson, "Is There Light at the End of the NGL Tunnel?" *Solid State Technology*, May 2000, p. 51
77. A.K. Bates *et al.*, "Review of Technology for 157-nm Lithography," *IBM J. Res. & Dev.*, Vol. **45**, No. 5, 2001.
78. A. Hand, "157-nm Optics Demand a Bag of Tricks," *Semiconductor International*, February 2002, p. 50.

Chapter 8

CHEMICAL MECHANICAL POLISHING (CMP)

As circuit features are scaled to submicron dimensions and additional levels are added to multilevel-interconnection schemes, the required degree of planarization is increased. In fact, the complexity of current microelectronic devices demands global planarity of the substrate after shallow-trench isolation and at all interconnect levels. Thus, such planarization needs to be implemented in both conductor and dielectric materials. In this chapter the technique of *chemical mechanical polishing* (the only known technique that provides global planarization for ULSI applications) is described. The discussion will begin by defining terms through which planarization is characterized, and then a brief recount of the evolution of planarization techniques that preceded CMP will be given. After concluding this introductory matter, the remainder of the chapter is devoted to the CMP process.

8.1 TERMINOLOGY & EVOLUTION OF PLANARIZATION PROCESSES FOR ICS

The term planarization is employed quite frequently (and often loosely) in other technical literature. Thus it is useful to define the term more completely. The example case used for this discussion is that of a dielectric layer deposited after Metal-1 is patterned (i.e., an *intermetal dielectric* [or IMD] film). In the case where no planarization exists (Fig. 8-1a), the step heights on the IMD surface closely approximate the step heights of the Metal-1-layer and the underlying topography. Furthermore, in this case the steps on the IMD surface generally also have steep slopes.

8.1.1 Defining the Degree of Planarization

The steps on the surface of the IMD can be made less severe through various planarization methods. The degree to which this can be successfully accomplished differs according to the technique used. The degree of planarization is classified according to the following qualitative planarization criteria:

1. The first degree of planarization (smoothing) involves a lessening of the step slopes at the IMD surface (Fig. 8-1b). The step heights in this case, however, are not significantly reduced in magnitude.

2. In the second degree of planarization (partial- or semi-planarization) the step heights are reduced (but not eliminated), and the slopes of the steps are also smoothed (Fig. 8-1c).

3. In the third degree of planarization (complete local-planarization), the steps at the surface of the IMD are completely eliminated wherever the spaces in the underlying topography are relatively close together (e.g., < 10-μm

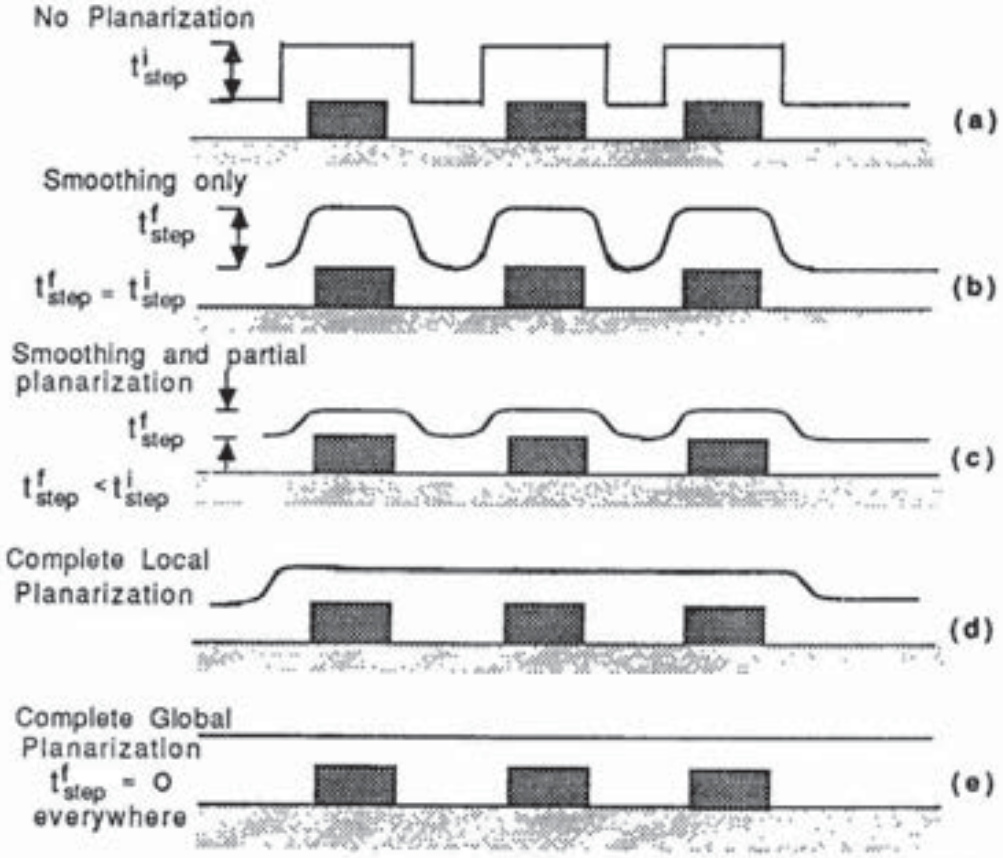


Fig. 8-1 (a) through (e) - Degrees of dielectric planarization; (f) SEM of 6-metal-level interconnect structure, with complete global-planarization realized at each level.

apart), but the steps at isolated, wide features still exhibit a step (Fig. 8-1d).

4. In the fourth degree of planarization (complete global-planarization), the surface is completely planarized over arbitrary topography (Fig. 8-1e).

Figure 8-1f shows a SEM of a 6-level-metal interconnect-system in which complete global-planarization has been realized in each of the metal levels (i.e., using CMP).

A quantitative measure of the step-height-reduction, referred to as the *planarization factor*, β, is given by[1]

$$\beta = 1 - (t^f_{step}/t^i_{step}) \quad \textbf{(8.1)}$$

where t^f_{step} and t^i_{step} are the final- and initial-step-heights, respectively. In complete-planarization cases, $\beta = 1$. If no planarization (or only smoothing) exists, then $\beta = 0$.

8.1.2 The Need for Dielectric Planarization

As the number of levels in an interconnect technology is increased, the stacking of additional layers on top of one another produces a more and more rugged topography. Consider, for example, a two-level-metal, single-poly CMOS technology. Assume the step-height of the semi-recessed field-oxide is 0.3-μm, and the thicknesses of the poly and the first and second metals are 0.4-, 0.5- and 1.0-μm, respectively. The maximum height of the steps on the wafer surface after each of these processes will be 0.3-, 0.7-, 1.2-, and 2.2-μm, respectively (Fig. 8-2). It is apparent that the surface of the wafer must be planarized in some fashion to prevent the topography roughness from growing with each level. Without such planarization, the microscopic canyons that result on the wafer surface from stacking of device features can lead to topography conditions that would eventually reduce the yield of circuits to unacceptably small values.

A nonplanar wafer topography results in three conditions that prevent the fabrication of reliable electrical connections in multilevel interconnect systems. These conditions are as follows:

1. Poor step-coverage of metal lines as they cross over the high and steep steps. A measure of how well a film maintains its nominal thickness is expressed by the ratio of minimum thickness of a film as it crosses a step, t_s, to the nominal thickness of the film on horizontal regions, t_n (Fig. 8-3a). This property is referred to as the *step-coverage* of the film, and is expressed as the percentage of the nominal thickness that occurs at the step:

$$\text{Step coverage (\%)} = (t_s/t_n) \times 100\% \quad \textbf{(8.2)}$$

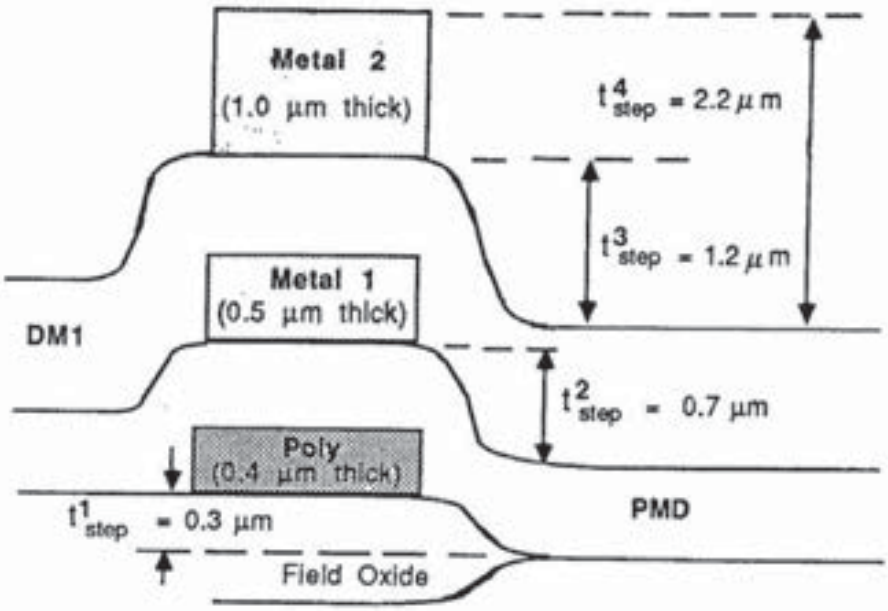


Fig. 8-2 As the number of levels in an interconnect technology is increased, the stacking of additional layers on top of one another produces steps of progressively greater heights.

Step coverage of 100% is ideal for a conductor, but in general the height of the step and the aspect-ratio of the features being covered impact the expected step-coverage. (The *aspect-ratio* is defined as the height-to-spacing ratio of two adjacent steps.) The greater the step-height or the aspect-ratio, the more difficult it is to attain coverage of the step without a corresponding thinning of the film that overlies it. Figure 8-3b shows an example of poor step-coverage. Hence, worse step-coverage is expected under these conditions. In addition to the step-height and aspect-ratio, step-coverage depends on two other topological-factors: the contour and slope of the step. In general, the smoother the contour of a step and the less vertical its slope, the better will be the step-coverage of an overlying film.

2. Metal stringers remain behind at the foot or sides of steep steps when anisotropic-etching is used. This condition is shown in Fig. 8-3c. In addition, the resist is typically eroded by the same gases that etch the metal. The resist must therefore be thick enough in all locations that it will not be completely removed before the etch-process is completed. If the dielectric surface has steps, in certain locations (i.e., over the upper corners of the steps in the dielectric, see Fig. 8-3d) there may not be sufficient resist thickness to survive during the time of overetch that must be used for removal of the stringers. The resist-erosion-rate during the overetch time

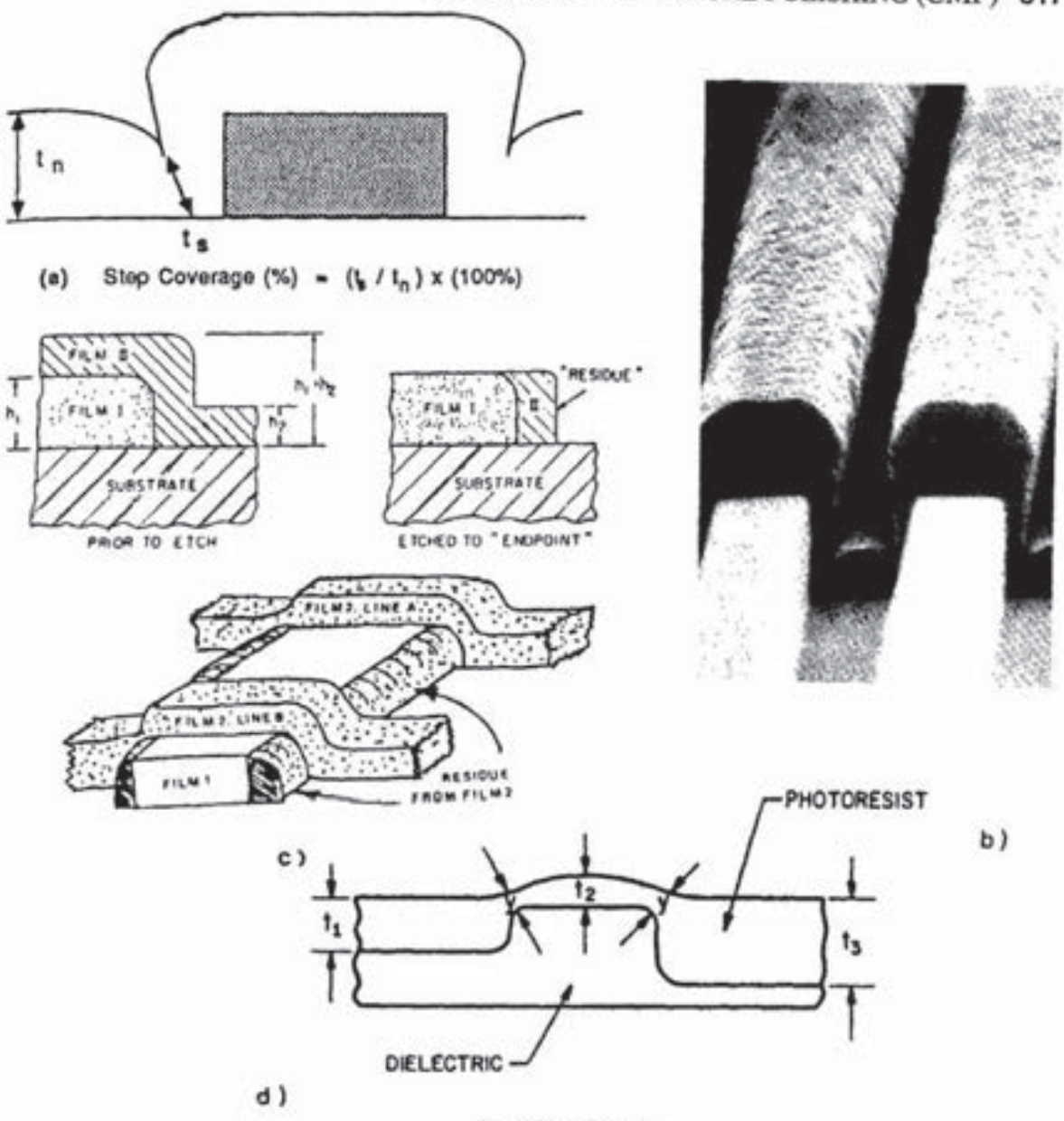


Fig. 8-3 a) Definition of step-coverage. b) SEM micrograph showing poor step-coverage. c) Metal stringers remain behind at the foot or sides of a steep step when anisotropic-etching is used. d) Problem of resist-thinning as film crosses underlying steps.

is also generally enhanced. Because most of the exposed-metal has been etched away, the reactant gases that were being consumed during etching of the metal are now present in greater concentrations. In cases where the resist is completely eroded at thin spots, the metal beneath the resist will be exposed to the reactive species and may be etched away to an unacceptable degree.

Even if the above two limitations could be overcome, the following obstacle would eventually force planarization to be implemented:

3. The depth-of-field limitations of submicron optical-lithography tools will require surfaces to be planar within ±0.5 μm. As a result, if steps larger than 0.5 μm exist on the surface of the IMD (or any other layer on the wafer), it will not be possible to pattern features in Metal-1 to the maximum resolution of the stepper. Thus, planarization will be mandatory if optical lithography is to be usable for fabricating ICs with such submicron feature sizes. (This topic is discussed in more detail in Vol.1, 2nd Ed., Chap. 13.)

Planarization of the intermetal dielectric layers is one of the chief approaches that has been taken to alleviate the problem of rough surface topography.

8.1.3 Design Rules Related to Intermetal Dielectric-Formation and Planarization Processes

The type of intermetal dielectric material, its thickness, and the method of deposition and planarization can all have an impact on the design rules that must be applied when laying out an integrated circuit with a particular multilevel-metal system. Here a two-level-metal system with an underlying poly-silicon level is considered. Assume that the topography beneath the polysilicon is completely planar. In this system, design rules dealing with the following conditions must be specified: **1**) minimum polysilicon-to-Metal-1 spacing, assuming no metal overlap of the poly; **2**) minimum distance between coincident edges when Metal-1 overlaps polysilicon; and **3**) minimum spacing of Metal-1-to-Metal-1.

If polysilicon and Metal-1 come too close together, a gap will exist between them that will lead either to void formation in the CVD oxide PMD layer following deposition (Fig. 8-4a) or to the formation of a deep, narrow crevice that cannot be covered by the next level of metal, even after some planarization procedures have been utilized (Fig. 8-4b).

In the second case (Fig. 8-4c), if polysilicon and metal edges are coincident, the total step height may be too high for adequate coverage by Metal-2 to be achieved. Figure 8-4d illustrates the case in which the problems of coincident edges and a narrow crevice are exacerbated by their simultaneous occurrence.

In the third case, the spacing between adjacent Metal-1 lines must be large enough to prevent void or crevice formation in the deposited dielectric film. The step that exists in the LOCOS field oxide must also be considered in real situations when such design rules are being formulated. One possible way to prevent the formation of voids, crevices, and excessively high steps, is to allow

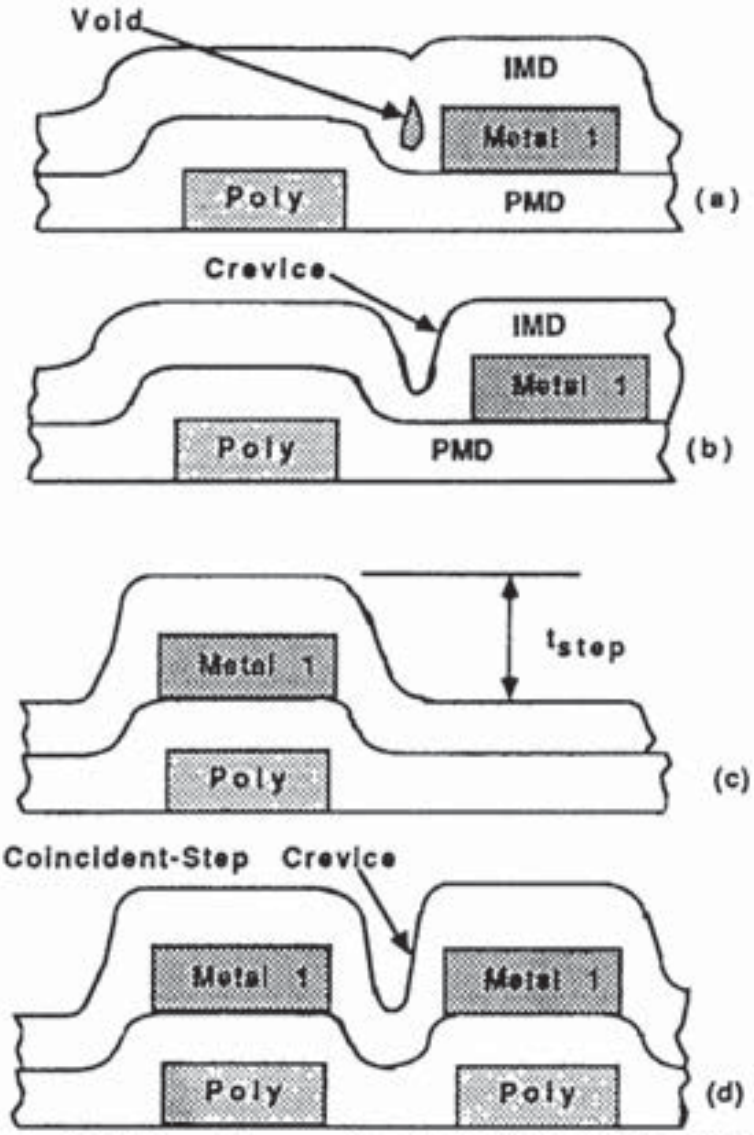


Fig. 8-4 Conditions associated with the relative spacing of polysilicon and Metal-1 that lead to layout design rules of these layers.

the spacings between Metal-1 and underlying features to be sufficiently large so any possible combination of misalignment cannot lead to the problems listed. However, this alternative is impractical, because it increases chip size and causes excessive complexity in layout design.

The trend in advanced IC designs today is toward the elimination of all such restrictive dielectric-layer-related design rules. The development focus is on dielectric planarization processes that allow all spaces between any adjacent or overlapping features (and all possible step heights), to be covered with a dielectric layer whose surface does not exhibit any voids, crevices, or excessively

high steps. Even though this is an actively pursued goal, it has not yet been attained in all processes and technologies. Various planarization approaches that have been developed along the path toward the condition in which no restrictive design rules need be invoked will be described.

8.1.4 Step-Height Reduction of Underlying Topography Alleviates the Need for Planarization

In two-level-metal processes it may be possible to avoid performing any planarization procedures with respect to the IMD if the topography underlying IMD is planarized by the process steps that precede IMD deposition. Although this approach may appear to be obsolete for ULSI circuits, several of the step-height-reduction techniques turn out to be useful for providing a planar surface under Metal-1. Such a condition will ultimately be required for three- and four-level-metal processes. The survey of planarization techniques of dielectric layers begins by considering this approach. In this section one older step-height reduction technique is described, namely *planarization through sacrificial-layer etchback.*

8.1.5 Planarization through Sacrificial-Layer Etchback

Planarization of CVD-interlevel-dielectric films can also be achieved using the *sacrificial-layer etchback technique* (Fig. 8-5). This method gained the most widespread acceptance in two-level-metal processes down to ~1-μm device tech-

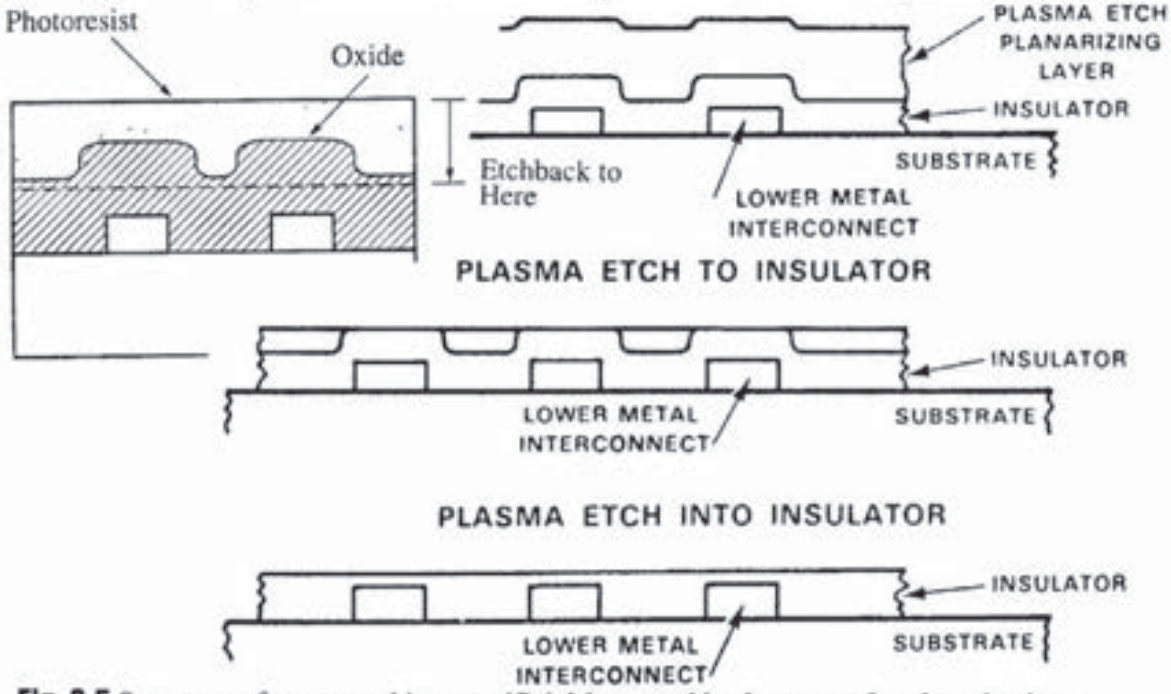


Fig. 8-5 Sequence of steps used in a sacrificial-layer etchback process for planarization.

nologies (and even in some reported three-level-metal bipolar processes). Using this approach, it is possible to achieve a high degree of planarization between steps that are ~2-10-μm apart. With more closely-spaced features, the technique runs into problems. For more widely spaced steps, planarization is less complete. The process is carried out by first depositing the CVD film that will serve as the interlevel dielectric (ILD) or intermetal dielectric (IMD). This layer is then coated with a film that will later be etched off (sacrificed). In most cases, such sacrificial-layers are photoresists, spun-on in liquid form. Upon being baked, these liquids become solid thin films that are thick enough so they provide a film surface that is close to being flat, as shown in Fig. 8-6. As long as the feature-size is smaller than about 10-μm, the degree of planarization can exceed 70% with most photoresists.

In the next step, the sacrificial-layer is first rapidly etched back in a plasma (typically, O_2, or O_2 mixed with CF_4) until the topmost regions of the dielectric layer are just exposed (Fig. 8-5b). The etch chemistry is then modified so the sacrificial-layer material and the dielectric are etched at approximately the same

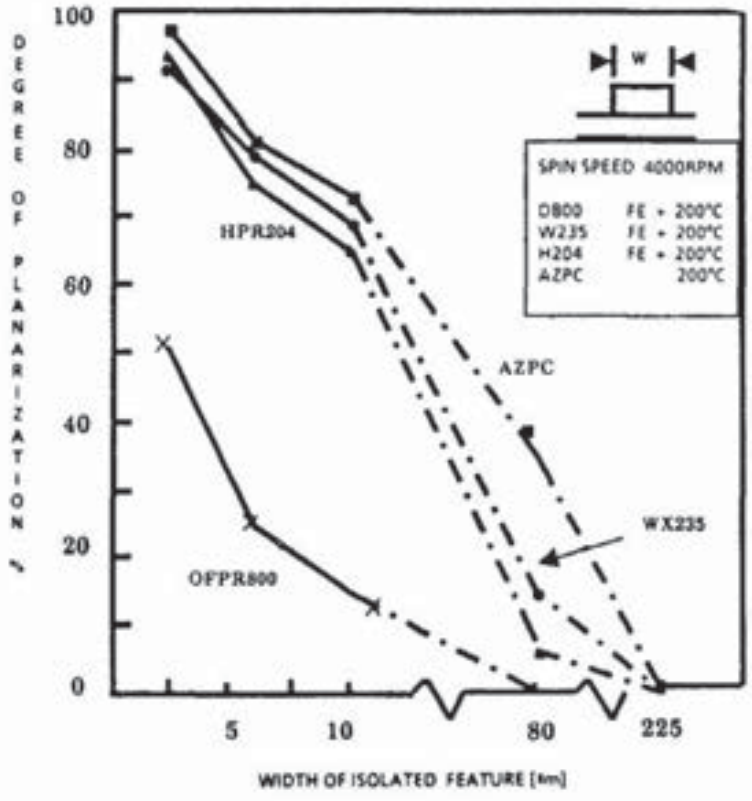


Fig. 8-6 Degree of planarization vs. various feature size for different polymers.[2] (© 1987 IEEE)

rate. The etch is continued under these conditions until all of the sacrificial layer has been etched away. At this point, the surface of the dielectric film is highly planarized since the profile of the sacrificial layer is transferred to the dielectric layer by the etchback procedure. The thickness of the dielectric film over underlying features (such as metal lines), may be thinner than desired after the etchback is completed. In fact, in some processes the etchback step is allowed to proceed until the Metal-1 lines are exposed. In either case, an additional layer of CVD dielectric is generally deposited in order to establish a minimum adequate thickness everywhere on the wafer surface.

8.1.5.1 Sacrificial-Etchback-Process Problems: To achieve the desired degree of planarity and final dielectric-film-thickness, very tight process control is necessary for such parameters as the magnitude of the etch-rate, the etch-rate uniformity across a wafer, end-point-detection, and oxide-thickness-uniformity under the resist layer. The resist etch-rate is sensitive to the cure cycle (especially if a high-temperature bake step [> 150°C] is used to improve surface planarity), and to chamber preconditioning. In batch processes, the etch-rate is also sensitive to the number of wafers present in the chamber. Such tight process-control of a complex process is hard to maintain on a production line.

The process also requires a multitude of steps, which implies a lengthy process, and thus low-throughput. Etchback is a difficult process to successfully implement in a batch-etch mode, resulting in an even lower-throughput. Finally, if the aspect-ratio between adjacent Metal-1 lines with vertical sidewalls exceeds 0.4, voids will form when most CVD-SiO_2 layers are deposited (Fig. 8-7). These voids can trap resist and moisture, which can subsequently be released if the voids are opened during the etchback process. Opened voids will also trap metal during the Metal-2 deposition-step. It will remain in the form of metal stringers following Metal-2 patterning. Thus, this process is not feasible for Metal-1 spacings of less than ~1.25-μm (unless a process is available for deposition of SiO_2 layers at low temperatures without formation of voids in high-aspect ratio spaces [e.g., PECVD TEOS films, O_3:TEOS films, or HDP CVD films]).

8.2 INTRODUCTION TO CHEMICAL MECHANICAL POLISHING (CMP)

As the number of interconnect layers in ICs has increased, the planarization of dielectric and metal layers has become more critical. In general, planarizing techniques such as the thermal flow of oxide layers, sacrificial-resist-etchback, and spin-on-glass, are inadequate for interconnect systems with more than three layers of metal. All of these techniques ultimately provide only a limited degree of smoothing and local planarization (i.e., over distances smaller than 10-μm).

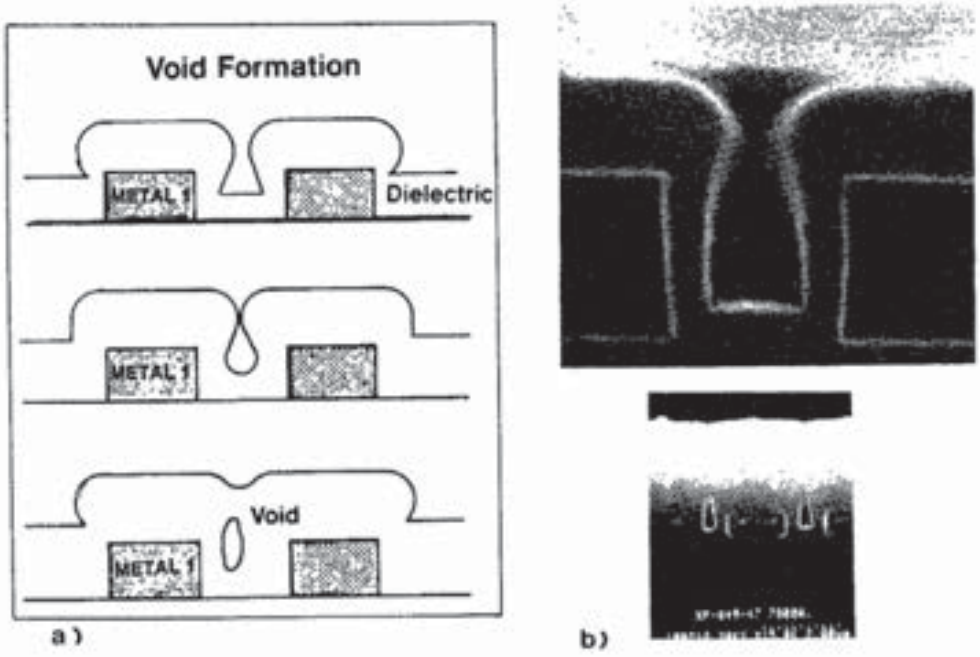


Fig. 8-7 (a) Diagram of void-formation due to cusping of the dielectric. (b) SEM of voids left between 1.5-μm-spaced Metal-1 lines in a 3.0-μm-thick PMD PECVD oxide.

They are not able to provide global planarization.[3] Global planarization of both dielectric and metal layers, however, is possible with chemical-mechanical-polishing (CMP), a process that was pioneered at IBM in the 1980s. In this technique, high elevation features are selectively removed (i.e., material from high elevation features is removed more rapidly than material at lower elevations), resulting in topography with improved planarity (see Fig. 8-10 [in which $h_{final} << h_{initial}$] and Fig. 8-10b). The degree of planarization is expressed by Eqs. 8-1 and 8-3. A step-height reduction of 90–95% is achievable using CMP, yielding values of β in the 0.90–0.95 range. CMP is also capable of planarization of distances up to several millimeters, far in excess of that possible with any other technique thus far developed.[4]

Mechanical polishing, assisted by chemical action, produces such selective material removal. The process is performed by mounting the wafer face down on a carrier. The carrier is then pressed against a moving platen containing a polishing pad. The carrier itself is rotated. An abrasive-containing aqueous slurry is dripped onto the table, and centrifugal force distributes the slurry across the pad. This forms a thin colloidal layer of slurry which saturates the pad (Fig. 8-8). The combination of mechanical effects and chemical reactions results in material removal from the surface of the wafer.

CMP offers three key advantages for fabricating sub-micron ICs. First, for logic devices (such as microprocessors and ASICs), CMP makes it possible to achieve high manufacturing yields and high device speeds with multilevel interconnects having up to eight levels of metal. By planarizing IC topography, CMP avoids the problem of metal thinning over steep topographies in multi-level interconnect structures. Second, memory devices (such as DRAMs) benefit from CMP even though they use interconnect systems with only two levels of metal. For memories, the flat wafer surface after CMP increases the depth-of-focus budget available for lithography (see Chap. 13, Vol. 1, 2nd Ed.). This allows memory-circuit designers to employ smaller critical dimensions, thus reducing chip sizes without decreasing yield. Smaller chip sizes provide a significant cost advantage in the competitive realm of DRAM manufacture. Third, CMP is paradoxically capable of reducing (rather than increasing) defect-density.

Because CMP has become an enabling technology for allowing ICs to be further scaled, the applications of CMP have become pervasive. For example, by 2000, stacked-capacitor DRAMs have upwards of 9 polishing steps, and microprocessors may incorporate as many as 16 polishing steps. As a result, the manufacturability aspects of CMP (namely, the overall cost per CMP step - expressed in terms of $/wafer/polish step) become as important as CMP yield.

In the remainder of this chapter the following topics will be discussed: **a**) the terminology used to characterize CMP processes; **b**) the history of CMP; **c**) the mechanisms involved in CMP of oxides and metals; **d**) CMP equipment and consumables; and **e**) the details of CMP for various materials; and **f**) the challenges facing CMP processing. The latter include: inadequate understanding of the polishing mechanisms, low throughput, inadequate reliability of the polishing tools, degradation of removal-rate with usage-time of a polishing pad,

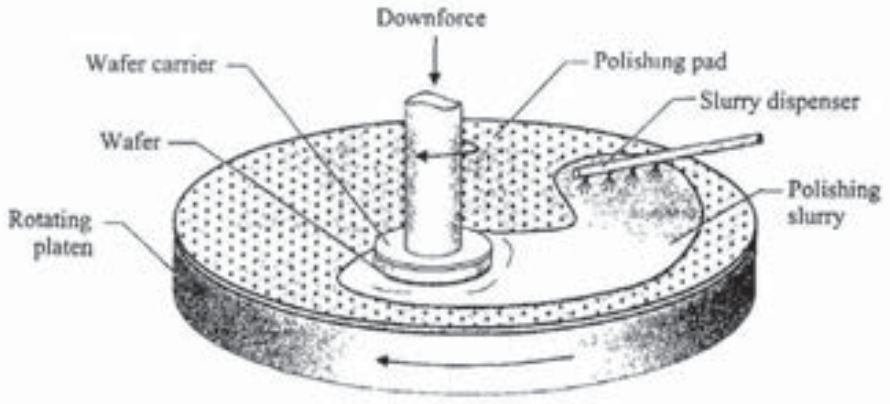


Fig. 8-8 Schematic of a rotary-table chemical-mechanical-polishing system.

non-uniformity of film-thickness-removal across the wafer, high cost of ownership, and high defect density.

8.3 THE TERMINOLOGY USED TO CHARACTERIZE CMP PROCESSES

Before beginning the discussion of the details of the CMP process, it is useful to define the metrics used to characterize the CMP process itself. This discussion is begun by listing the goals desired in a CMP process. They are:

1. The CMP process should have the ability to selectively reduce the step heights of features in the film being polished (i.e., a high degree-of-planarization must be possible);
2. Good planarity across a die (highest spot to lowest spot should be less than 150-nm);
3. Good film thickness uniformity within the wafer (±2–5%, 1σ);
4. High removal rates (100–140-nm/min), for high throughput and low cost-of-ownership;
5. Uniform removal rates from wafer-to-wafer;
6. Good selectivity and/or good endpoint technology to avoid under-polishing or over-polishing;
7. The CMP process should not add excessive defects (e.g., for 0.18-μm CMOS, fewer than 0.125-particles/cm^2 with a size larger than 0.06-μm are allowed at each mask level);
8. The overall cost of a CMP process should be as low as possible (expressed in $/wafer for each CMP process which is conducted).

While it is useful to list these goals in the qualitative manner as given above, the metrics for characterizing a CMP process can also be expressed in quantifiable form. These quantifiable forms are presented next.

8.3.1 CMP Removal-Rate (RR)

The *removal-rate* (RR) describes how rapidly a CMP process removes material from a non-patterned wafer (Fig. 8-9). This quantity is measured by performing a CMP process on a planarized, unpatterned wafer for a specific time, and determining how much material has been polished away. Generally speaking, the RR should be high so that the CMP process can be completed quickly enough to allow a satisfactory throughput. The RR is quantitatively expressed as follows:

$$RR = \Delta T/t \tag{8.3}$$

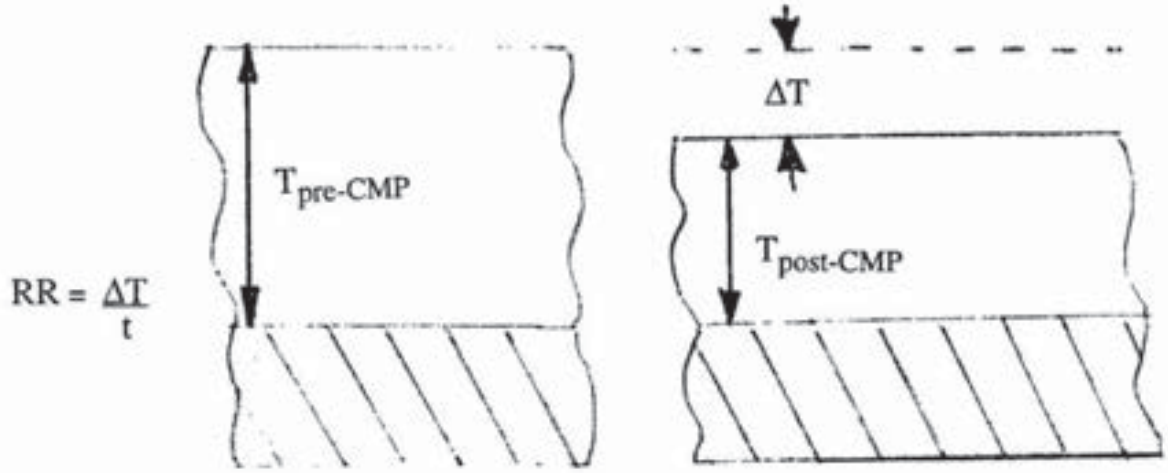


Fig. 8-9 Definition of *removal-rate* (RR).

where ΔT is the thickness of material removed in time t. RR is most frequently expressed in units of Å/min.

8.3.2 The Degree of Planarization (DOP)

In the ideal case, after CMP the surface should be completely planarized (see Fig. 8-1). The quantity that describes how well the CMP process meets this ideal is the *degree of planarization* (%DOP). The %DOP (see Fig. 8-10a) is quantitatively expressed as:

$$\%DOP = [(H_d^i - H_d^f)/H_d^i] \times 100\% \quad \textbf{(8.4)}$$

where H_d^i is the height of feature being polished prior to carrying out the CMP step, and H_d^f is the height of the feature after the completion of the polishing step. Note that when the value of the %DOP is 100% this indicates that the feature has been completely planarized. Figure 8-11b shows a wafer surface on which the %DOP is 100% (complete global-planarization). Figure 8-11a shows a wafer surface on which the CMP process has not produced complete planarization (%DOP < 100%).

8.3.3 Within-Die Non-Uniformity (WIDNU)

The rate at which material is removed by CMP within a die may not be uniform. Such *within-die non-uniformity* (WIDNU) is quantitatively defined as the *standard deviation of the film thickness within a die*, σ_{WID}, divided by the *average value* of all "*n*" film-thickness measurements within the die, $\bar{x}_{WID}$:

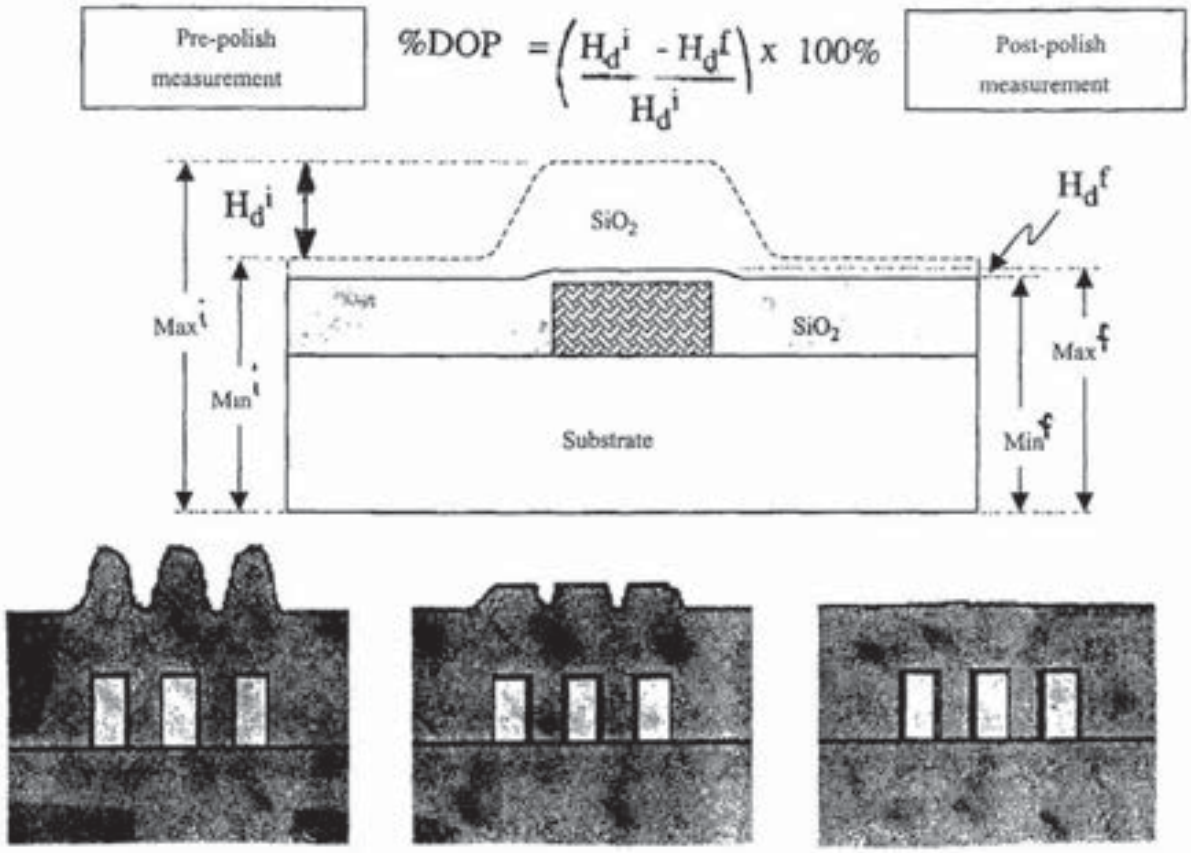


Fig. 8-10 Definition of *degree of planarization* (DOP).

$$\text{WIDNU} = \sigma_{\text{WID}} / \bar{x}_{\text{WID}} \qquad \textbf{(8.5)}$$

Note that σ_{WID} is the standard deviation of measurements within the die of some specific thickness, depending on the type of thickness non-uniformity being specified (e.g., the thickness variation of the oxide in the field between metal lines in an otherwise planarized oxide surface, see Fig. 8-11), and it is expressed quantitatively by:

$$\sigma_{\text{WID}} = \sqrt{[(x_1 - x_{\text{WID}})^2 + \ \ldots + (x_n - x_{\text{WID}})^2]/(n-1)} \qquad \textbf{(8.6)}$$

where $x_1 \ \ldots \ x_n$ are particular measurements on a die. Two examples of effects that cause WIDNU are shown in Figs. 8-12 and 8-13, respectively.

In Fig. 8-12c the step heights after CMP have not been reduced by the same amount within a die. Specifically, narrow, isolated-features tend to be planarized to a greater degree than wide, isolated-features, or narrow, densely-spaced-features. This form of WIDNU is termed *pattern-dependent-planarization.* The causes of such pattern-dependent-planarization are described in Sect. 8.5.3.2.

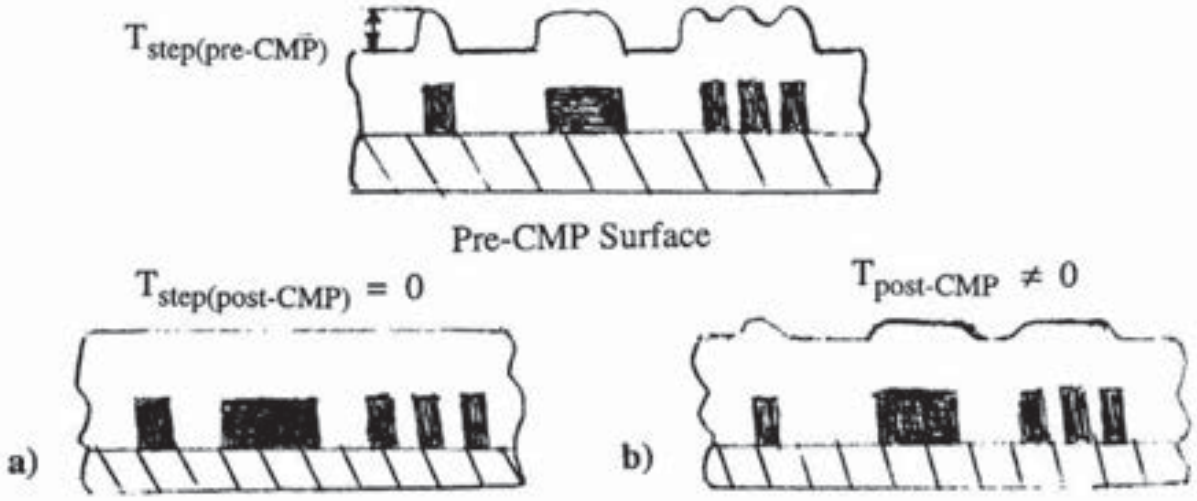


Fig. 8-11 a) Surface on which complete global-planarization has been achieved; b) Surface on which complete global-planarization has not been achieved.

In Fig. 8-13c, although the step heights of the oxide within a die (as shown in Fig. 8-13a) have been completely planarized, some film has been eroded in regions between the steps to a degree that the oxide surface is not planar throughout the die. This causes the distance from the top of the oxide to the planarized substrate surface below the oxide to differ within the die. This form of WIDNU is known as *dishing*, and the effects that cause such dishing are described later.

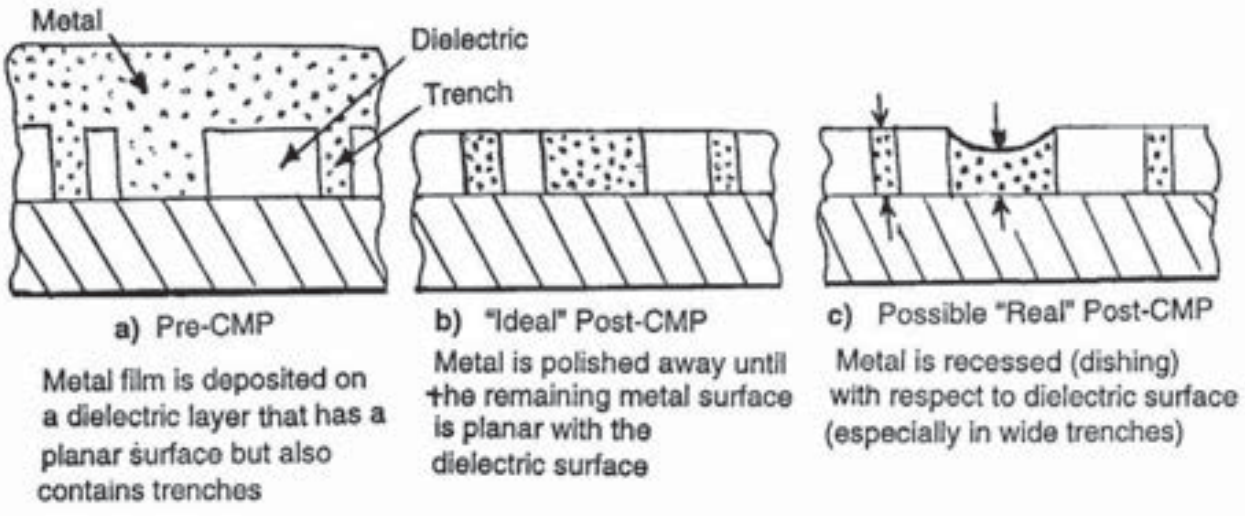


Fig. 8-12 Effect #1 that causes *within-die non-uniformity* (WIDNU).

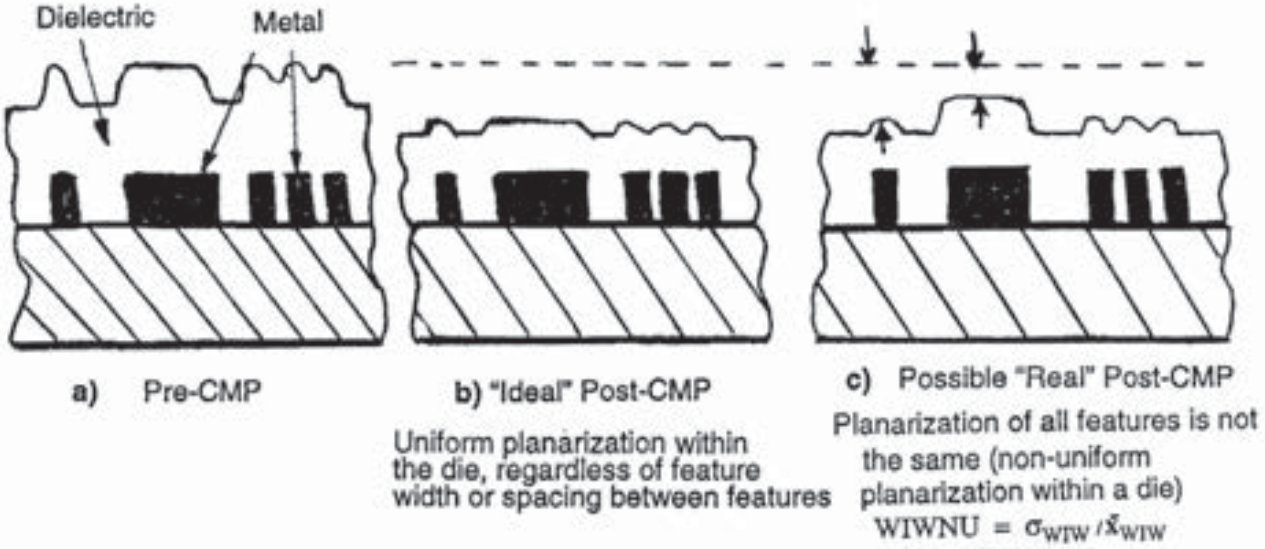


Fig. 8-13 Effect #2 that causes *within-die non-uniformity*, WIDNU (dishing).

8.3.4 Within-Wafer Non-Uniformity (WIWNU)

The *within-wafer non-uniformity* (WIWNU) is defined as the *standard deviation of the film thickness within a wafer*, σ_{WIW}, divided by the *average value* of all "*n*" film thickness measurements on that same wafer, $\bar{x}_{WIW}$.

$$WIWNU = \sigma_{WIW} / \bar{x}_{WIW} \tag{8.7}$$

Note that σ_{WIW} is the standard deviation of measurements within the wafer of some specific thickness, depending on the type of thickness non-uniformity being specified (e.g., the thickness variation of the distance between the otherwise planarized surface of the oxide and the surface of the planarized substrate below, see Fig. 8-14), and it is expressed quantitatively by:

$$\sigma_{WIW} = \sqrt{[(x_1 - x_{WIW})^2 + \ \ldots + (x_n - x_{WIW})^2]/(n-1)} \tag{8.8}$$

where $x_1 \ \ldots \ x_n$ are particular measurements on a wafer. Two examples of WIWNU are shown in Fig, 8-15.

In Fig. 8-15a the surface of the oxide after CMP is shown to be globally planarized across the entire wafer. However, the distance from the surface of the oxide to the planarized surface of the substrate below is non-uniform across the wafer. An example of two effects that can lead to such WIWNU are: **1**) If the thickness of the substrate (wafer) itself is non-uniform, and the CMP process is performed with a completely rigid pad, this will cause WIWNU (see Sect.

$$\bar{x}_{WIW} = (x_1 + x_2 + x_3 + x_4)/4 \qquad WIWNU = \frac{\sigma_{WIW}}{\bar{x}_{WIW}}$$

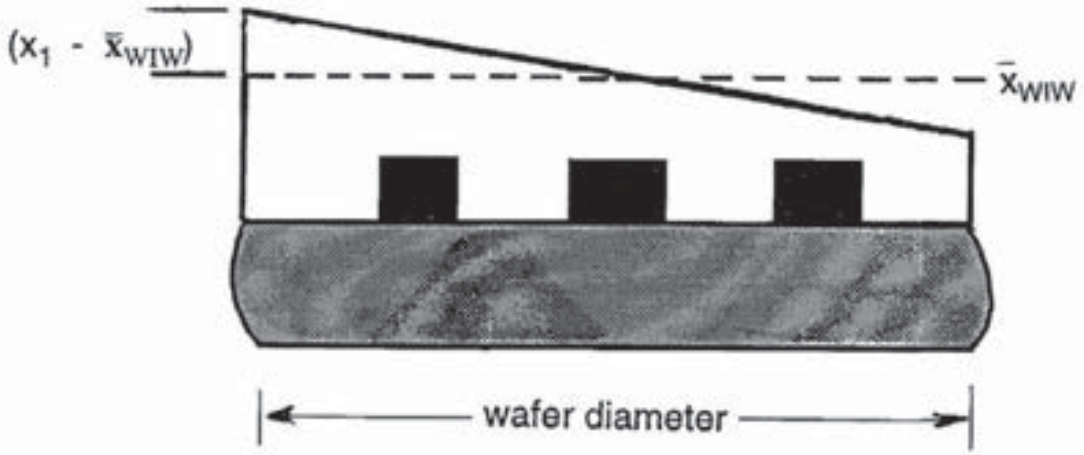


Fig. 8-14 Definition of *within-wafer non-uniformity* (WIWNU).

8.12.2); and **2**) In rotary CMP polishing tools, WIWNU can be caused by the relationship between the angular rotation velocities of the wafer carrier and the polishing pad platen. This topic is discussed in Sect. 8.5.

Another effect that gives rise to a special type of WIWNU is shown in Fig. 8-15b. This figure shows that the wafer surface after CMP has an edge bead of material, caused by a smaller removal rate of material by CMP near the edge of the wafer. This topic is described further in Sect. 8.6.2.1.

8.3.5 Wafer-to-Wafer Non-Uniformity (WTWNU)

The *wafer-to-wafer non-uniformity* (WTWNU)) is defined as the *standard deviation of the film thickness from wafer-to-wafer*, σ_{WTW}, divided by the

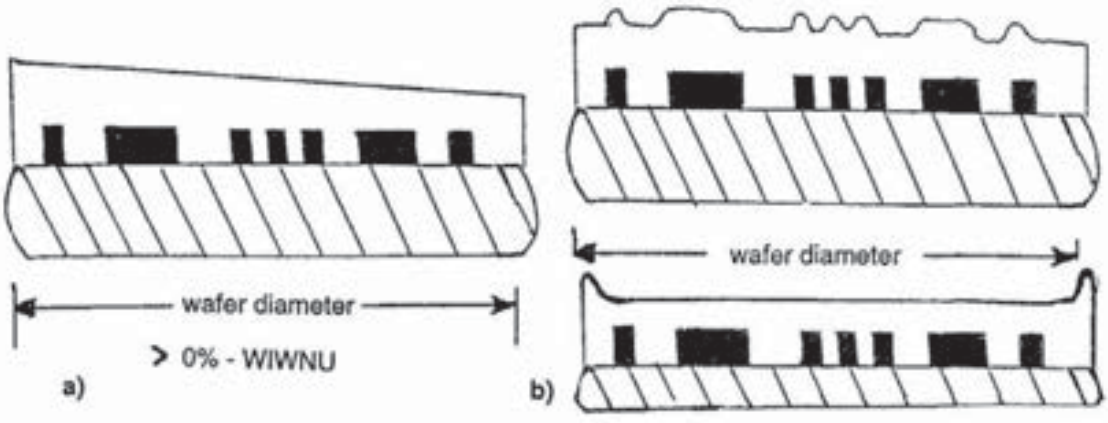


Fig. 8-15 a) Example of one type of WIWNU; b) Example of a second type of WIWNU caused by the *edge-bead-effect*.

average value of all the measurements of the film thickness measurements taken over all n wafers, $\bar{x}_{WTW}$.

$$WTWNU = \sigma_{WTW} / \bar{x}_{WTW} \quad \textbf{(8.9)}$$

Note that σ_{WTW} is the standard deviation of measurements from wafer-to-wafer for "n" wafers of some specific thickness, depending on the type of thickness non-uniformity being specified (e.g., the thickness variation of the distance between the otherwise planarized surface of the oxide and the surface of the planarized substrate below, see Fig. 8-16), and it is expressed quantitatively by:

$$\sigma_{WTW} = \sqrt{[(x_1 - x_{WTW})^2 + \ldots + (x_n - x_{WTW})^2]/(n-1)} \quad \textbf{(8.10)}$$

where $x_1 \ldots x_n$ are particular measurements on a wafer. An example of WTWNU is shown in Fig. 8-16. Two of the many causes of WTWNU include pad-aging (see Sect. 8.6.7), and slurry inconsistency from batch-to-batch (see Sect 8.6.3).

8.3.6 Efficiency of Planarization (EOP)

When an oxide is deposited over a metal line, enough oxide over the metal line must exist after the CMP step to minimize electrical cross-talk and parasitic capacitance between metal layers. This oxide thickness is determined by the IC designer. The CMP process-engineer must set up the CMP process so that after CMP is completed the oxide minimally meets the designers' requirements. The *efficiency of planarization* of a CMP process has an impact on the process integration choices that must be made to ensure that the minimum oxide thick-

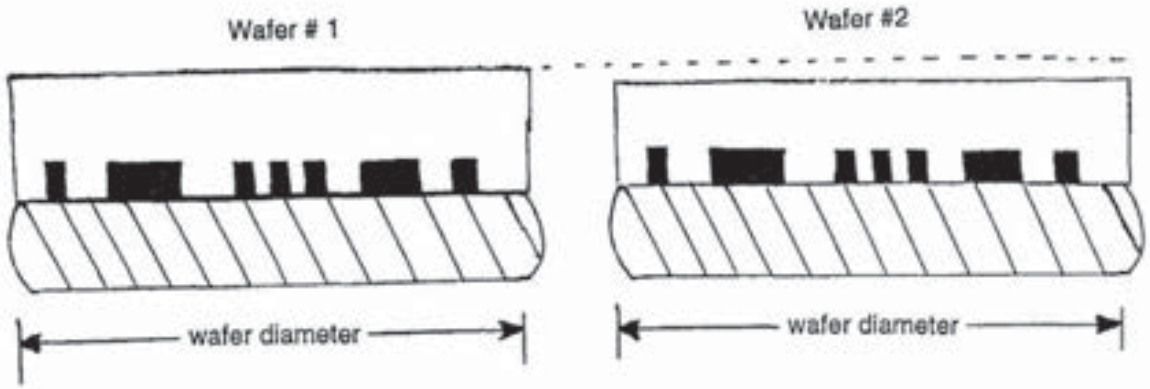


Fig. 8-16 Definition of *wafer-to-wafer non-uniformity* (WTWNU).

ness remains after CMP. The *efficiency of planarization* (%EOP) is expressed as:

$$\%EOP = [(T_r - T_d)/ T_r] \times 100\% \quad \textbf{(8.11a)}$$

where T_r is the dielectric thickness removed by the CMP process on top of a feature (e.g., a metal line) and T_d is the oxide-thickness-loss at the center of a wide space between metal lines (see Fig. 8-17). The ideal case would be a CMP process with an %EOP = 100%. If this occurs, that means that complete planarization is achieved without the erosion of any oxide in the center of the wide space between metal lines. However, the %EOP is generally not 100%, and thus some erosion of the oxide between metal lines usually occurs.

It is the task of the CMP process engineer to determine how much total oxide must be deposited prior to the CMP step to ensure that the oxide on the metal lines is at least as thick as that specified by the IC designer. This thickness quantity is normally determined during tool setup, but it is useful to have an expression which allows the CMP engineer a way to estimate the correct value prior to performing the experimental task itself.

To allow the CMP engineer to calculate this thickness value more easily, the %EOP expression given by Eq. 8-11a is rewritten using different parameters. That is, Eq. 8-11a provides an expression for %EOP that makes it easier to visualize the concept of %EOP, but using it makes it cumbersome to calculate the required oxide deposition-thickness. Therefore, another expression is used instead of Eq. 8.11a for the deposited-oxide-thickness calculation. This alternate relationship is given as Eq. 8-11b (see Figs. 8-17c and 8-17d)

$$\%EOP = [H_{metal} /(T_{total} - T_{cov})] \times 100\% \quad \textbf{(8.11b)}$$

where H_{metal} is the height of the metal line being covered by the oxide, T_{total} is the total amount of the oxide film deposited over the metal line prior to CMP, and T_{cov} is the target amount of oxide that should remain over the metal line (as specified by the IC designer). Equation 8-11b is then rearranged to allow T_{total} to be determined:

$$T_{total} = (H_{metal}/EOP) + T_{cov} \quad \textbf{(8.12)}$$

As an example, if the %EOP is 50%, and H_{metal} = 0.5-μm, and T_{cov} is specified as 0.5-μm, then using Eq. 8-12 (with EOP = %EOP/100% = 50%/100% = 0.5), then the amount of oxide that would have to be deposited prior to the CMP process that would result in 0.5-μm of oxide over the metal lines after CMP would be T_{total} = 1.5-μm.

The least-expensive process involves depositing the smallest thickness of oxide (and this would occur if %EOP was 100%, corresponding to a T_{total} = 1.0-

$$\%EOP = [H_{metal} / (T_{total} - T_{cov})] \times 100\%$$

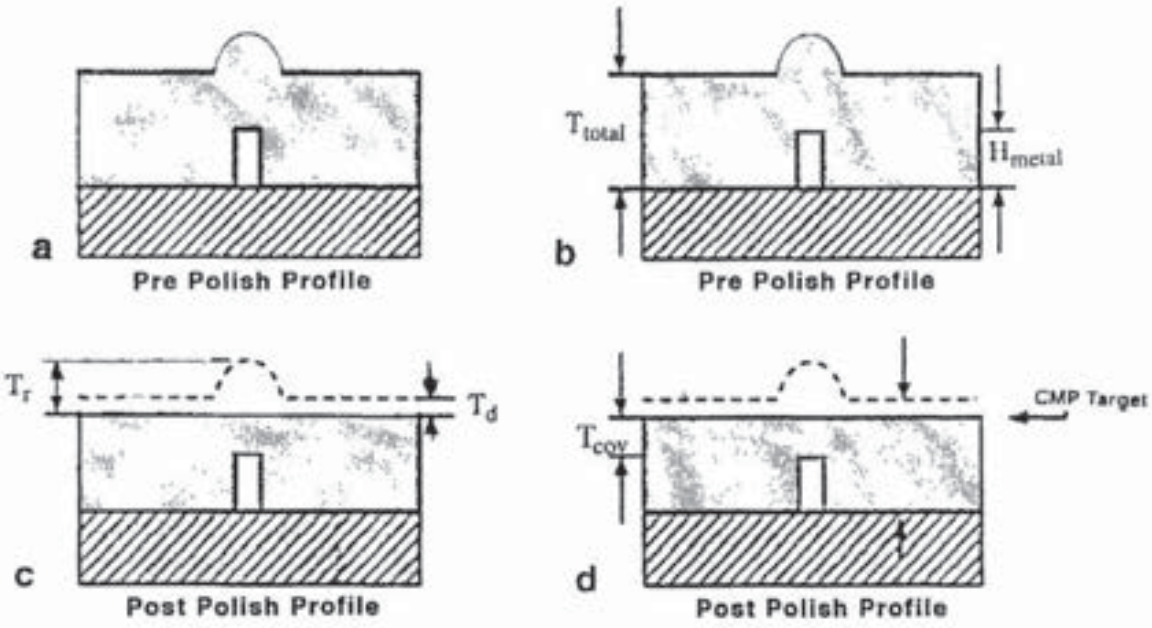


Fig. 8-17 The basic CMP process is one of reducing film-thickness selectivity over the wafer, ending with improved planarity. Here that metric is defined as *efficiency of planarization* (EOP).

μm). Thus, it can be seen that if the %EOP can be increased, the overall IC fabrication cost associated with forming the planarized oxide film over the metal steps will be reduced. For example, if %EOP could be increased to 80% (from 50%), Eq. 8-12 would predict that the T_{total} = 1.1-μm.

8.4 THE HISTORY OF CMP

The early history of CMP was chronicled by Fury.[5] He recounts that the process of chemical-mechanical-polishing originated at IBM in 1983. Research and development on CMP was exclusively performed within IBM until 1988. During those five years IBM closely guarded all information about the process. By the late-1980s several factors changed their secretive attitude. First, Intel microprocessors were selected for IBM's PC product line, and IBM decided to share some CMP data to help improve Intel's interconnect technology. Second, at that time the DRAM market was dominated by Japanese manufacturing companies. IBM was wary of having to rely entirely on foreign suppliers for such a critical component of its computer systems. Hence, they elected to help

strengthen a domestic DRAM supplier (Micron Technology, Inc.) by teaching them about the CMP process. Information about CMP was included in the process for the 4-MB DRAM that IBM sold to Micron Technology in the mid-1980s. In 1988, IBM agreed to disclose to SEMATECH its process scheme for multi-level interconnects, although they did not give all the details on CMP. They did, however, provide one of their CMP polishing tools (a Si-wafer polishing-tool sold by Westech, and marketed later as the IPEC Model 372, shown in Fig. 8-18). Motorola was introduced to CMP through the Power-PC microprocessor project, a joint venture between IBM, Motorola, and Apple Computer. In 1989 IBM published the first paper which publicly discussed the CMP of oxide films.[6] These events initiated the rapid proliferation of CMP throughout the IC industry.

The spread of CMP was hastened by other events, chiefly the downsizing of IBM from 1990–94. Many engineers with expertise in CMP left IBM and went to work for other IC companies who were seeking to implement the technology. The effort at SEMATECH involved improving the CMP tool-performance and reliability. Their work with Westech tools led to the development of the early popular CMP-tools, sold later as the IPEC 372 and 472 models. (Westech was acquired by IPEC.)

Although early development of CMP was centered in the U.S., worldwide efforts soon got underway. Polishing tools were initially offered by companies that had been building machines for polishing silicon wafers (Speedfam and Strasbaugh, as well as IPEC [Westech]). Overseas suppliers soon joined them (e.g., Ebara and Cybeq from Japan, and Wolters from Europe). Eventually the

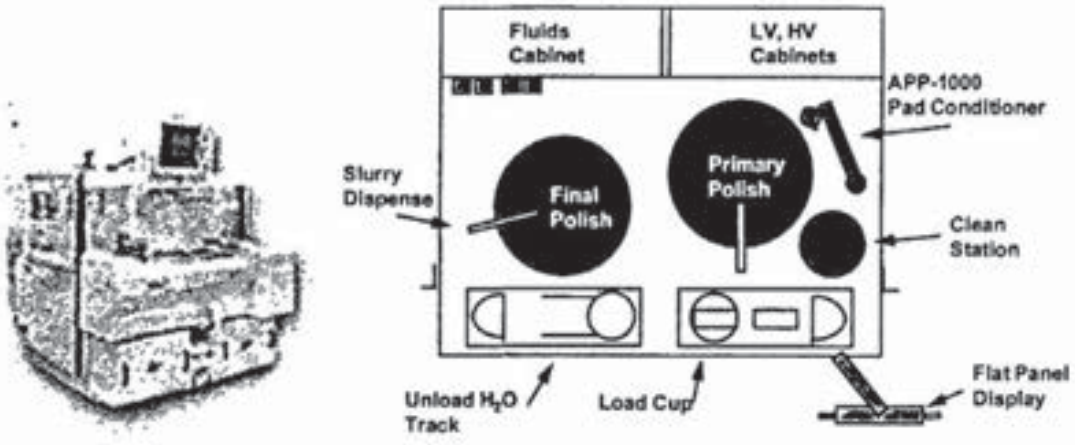


Fig. 8-18 The Speedfam 372 CMP tool.

larger semiconductor equipment vendors entered the market. Applied Materials introduced its tool in 1996, and Lam Research acquired On-Trak Systems, another commercial CMP vendor. The large number of vendors gave rise to some consolidation as well (an example of which was the merger of Speedfam and IPEC in 1999).

CMP was the fastest growing IC process technology of the 1990s, and it was also the most rapidly growing segment of the semiconductor process equipment market in that time frame. From 1995 to 1997 sales tripled, reaching $520 million in 1997. Sales were expected to exceed $1 billion by 2000.[7] The number of CMP polishing steps per year was also estimated to reach 125 million by that date. The first application of CMP was the planarization of dielectric films (mainly CVD oxide), but the number of applications has increased. Currently it is being used to planarize metal films (including tungsten, aluminum, and copper), as well as oxides and low-*k* dielectric films. By the end of the century, CMP had evolved from being a cleanroom-engineers' nightmare, to a critical process technology without which the production of ULSI chips with multilevel-interconnects might not be possible.

8.5 MODELING THE MECHANISMS OF CMP

The mechanisms of CMP are quite complex, and differ for each type of film being polished. Many variables also impact the removal rate, including: film type, pressure (force applied to the polishing pad), velocity, temperature, slurry feed rate, type of polishing motion, slurry particle size, pad hardness, and pad conditioning. The pressing need for production-worthy CMP capabilities forced early researchers to ignore the complexity of the process and go straight to an empirical trial-and-error approach. But, to logically optimize a CMP process, a more fundamental understanding of the mechanical and chemical forces at work is needed. While a complete understanding of CMP mechanisms has not yet been achieved, some insight has been gained. A brief introduction as to what is known about CMP mechanisms is presented here, divided into three categories: **1**) the mechanical aspects of material removal by CMP (Preston's Equation); **2**) the chemical and mechanical effects that cause planarization of oxides by CMP; and **3**) the chemical and mechanical effects that cause planarization of metals by CMP.

8.5.1 The Mechanical Aspects of Silicon Dioxide Removal by CMP: (Preston's Law)

The CMP process for planarizing films in ULSI devices evolved from the technology of chemical-mechanical-polishing used to produce silicon wafers. That is, CMP in the Si wafering process is used to: **1**) remove residual sawing and grinding damage that occurs when single-crystal silicon ingots are sliced into wafers; and **2**) enhance the flatness (planarity) of the bare, starting-wafer surface.

However, the CMP process used in the Si wafering procedure differs significantly in two ways from the oxide-CMP process used to planarize oxide-layers a ULSI IC:

1. The oxide-CMP process is much more exacting than the Si-wafering CMP process. That is, the thickness of material removed in the CMP step of wafering is on the order of 10-μm ± 0.1 μm. However, the thickness of material removed in the oxide-CMP processes is only on the order of 0.5-μm ±0.01 μm (i.e., about 5% of the thickness of the material removed in Si wafer polishing). Thus, the degree of control in oxide-CMP processes is much higher than that needed for Si wafer polishing.

2. When a *bare Si wafer is being polished*, the process involves oxidizing the Si surface with a chemical in the slurry solution, and then the slurry particles abrade away this oxide. That is, the CMP of Si wafers is carried out with a slurry containing fine slurry particles dispersed in a solution that *oxidizes the bare-Si wafer.* This is the *chemical component* of the CMP process. The abrasion of this oxide by the slurry-particles is the *mechanical component* of the process. The abrasive process exposes the Si beneath the oxide to the chemical in the slurry, which re-oxidizes the exposed Si, and so on. In fact, the abrasion of the oxide performs two functions: **a**) The polishing process is continued long enough so that the damaged region near the Si surface from sawing and grinding is eventually consumed by the oxidation and abrasion sequence; and **b**) Since the abrasion occurs more rapidly on "high" spots on the wafer surface, these "high" spots are eroded more rapidly than the "low" areas. Thus, the wafer-surface becomes flatter (more planar).

However, when an oxide is being polished in a ULSI IC (a process that is carried out in such applications as the CMP of the trench-oxide in shallow-trench-isolation structures (STI), and the CMP-step used to planarize the inter-level dielectric [ILD] and inter-metal dielectric [IMD] layers), the material to be polished *is already an oxide.* Hence, the

mechanism employed to polish the Si wafer (i.e., chemically oxidizing the Si surface, and mechanically abrading it away), is not an operative mechanism in oxide-CMP. A discussion of what are believed to be the phenomena that govern the removal of oxide films in oxide-CMP follows.

In 1997 Tomozawa stated that, like optical-glass polishing, much of the oxide CMP process was still an art rather than a technology. To qualify as a technology, the basic mechanisms of a process must be understood, allowing it to be logically optimized. However. at the time Tomozawa made his comment, the basic mechanisms of oxide CMP were not yet well understood.[8] The descriptions provided in this section briefly summarize what is known about the phenomena associated with oxide-CMP. Since silicon oxides used in IC fabrication are a form of glass, oxide-CMP is a specialized form of the glass polishing technology used to produce optical lenses. Thus, some aspects of oxide-CMP can be understood from the long history of glass polishing (which generally have been derived from trial-and-error, rather than from a fundamental understanding of the process). Optical-glass and SiO_2-films-on-Si-wafers are polished with a slurry consisting of abrasive silica particles (whose hardness is about equal to that of the oxide) suspended in an aqueous alkaline solution (typically KOH or NH_4OH, diluted in DI-water).

An empirical relationship that models the optical-glass-polishing removal-rate, RR, of flat-glass regions (defined as the thickness of glass removed per unit time), was developed by Preston, and is known as *Preston's equation* or *Preston's Law*.[9] This law states that the RR is proportional to the applied-pressure P and linear-velocity v at which a polishing pad moves relative to the optical-glass specimen, or:

$$\mathrm{RR} = K_p P v \qquad \textbf{(8.13)}$$

where K_p (*Preston's constant*) is a proportionality constant that is independent of pressure and velocity. The same law appears to apply to the rate of removal of oxide by CMP. (Typical values of P and v in CMP processes are 7-psi, and 30-40-ft/min, respectively.)

Note that Preston's equation describes the relationship between the RR of oxide and *mechanical effects*. (That is, mechanical effects are those caused the application of physical forces on a substance, namely in Preston's equation, the pressure exerted on the oxide and the relative velocity between the glass and the polishing-pad.) The Preston equation predicts that oxide-RR is linearly proportional to P and v, but gives no insight into other phenomena (i.e., such physical

and chemical factors as the glass-temperature, the chemical nature of the polishing medium, the hardness of the glass, etc). These effects are all put into the Preston constant, K_p, which contains the influence of such phenomena as glass hardness, Young's modulus, the polishing slurry, and the composition and hardness of the polishing pads. Despite the fact that Preston's equation sheds no light *per se*, on how these other phenomena impact RR, it is still a model that provides useful predictability concerning oxide-CMP. Hence, it is informative to discuss some further issues related to this equation.

The most important criticism of the Preston equation is that it was formulated empirically rather from first principles (i.e., it was not mathematically derived from a physical model of the glass-polishing process). Several authors have proposed equations from first principles that predict the dependency of oxide RR on pressure and velocity, including Brown and Cook (1984),[10] Liu *et al.* (1995),[11] Runnels and Eyman (1994),[12] Tseng (1997),[13] and Zhao and Shi (1999).[14] A critical review of the evolution of these models is found in Shinn *et al.*[15]

Many of the models (and supporting experimental evidence) indicate that the dependence of oxide-RR in CMP on pressure and velocity is not a simple linear relationship. Recent work has proposed multiple variations of Preston's equation of the form:

$$\mathrm{RR} = K_p\, v^{\beta}\, p^{\alpha} \tag{8.14}$$

Where α and β are adjustable parameters dependent on the system to be polished. The values of these coefficients must be experimentally determined.

In the model of oxide RR of Zhao and Shi (which is the first model that also includes the effect of pad and wafer hardness), the RR demonstrates a sub-linear dependence on pressure given by:

$$\mathrm{RR} = K_p\, v\, P^{2/3} \tag{8.15}$$

Figure 8-19 shows experimental data demonstrating the dependence of RR for oxide films.[16] Zhao and Shi also found that the pressure dependence of RR significantly influences planarization efficiency and pattern-density dependence. For instance, not all slurry particles in contact with a wafer necessarily remove material efficiently. A particle rolling against the surface removes little material, yet a particle held firmly by the pad slides against the wafer surface, changing the surface structurally and chemically. Sliding and removing requires a minimum local pressure, giving a revised polishing rate of:

$$\mathrm{RR} = K_p\, v\, (P^{2/3} - P_{\mathrm{th}}{}^{2/3}) \qquad \text{when } (P \geq P_{\mathrm{th}}) \tag{8.16}$$

and

$$\mathrm{RR} = 0 \qquad \text{when } (P \leq P_{\mathrm{th}}) \tag{8.17}$$

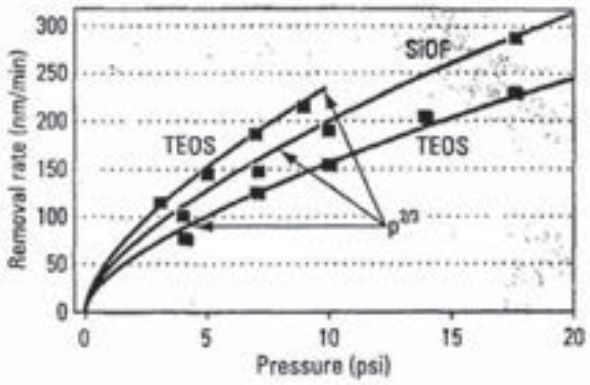


Fig. 8-19 Experimental data correlates well with the $P^{2/3}$ dependence of removal-rate on applied pressure.[16]

where P_{th} is the threshold pressure. In addition, they found that planarization efficiency also improves with P_{th} (Fig. 8-20), requiring less material removal to achieve a given planarization result.[16]

8.5.1.1 Model that Treats WIWNU Effects in Rotary-CMP Tools Based on Preston's Law: One of the limitations of CMP is that within-wafer-non-uniformities (WIWNU) occur, as defined in Sect. 8.3.4. The Preston equation states that the oxide-RR is linearly proportional to the *linear velocity* of the pad with respect to the oxide-surface being polished. Thus, one way to reduce such WIWNU is to ensure that the same linear velocity is experienced by all areas of the wafer. In rotary-CMP tools, the wafer carrier rotates, as does the platen that holds the polishing pad. If the pad did not rotate, and only the carrier rotated about its axis, the oxide-RR at center of the wafer would drop to zero, since there would be no relative motion between the wafer and the pad at that point. In addition, the RR would increase radially from the center to the edge of the wafer.

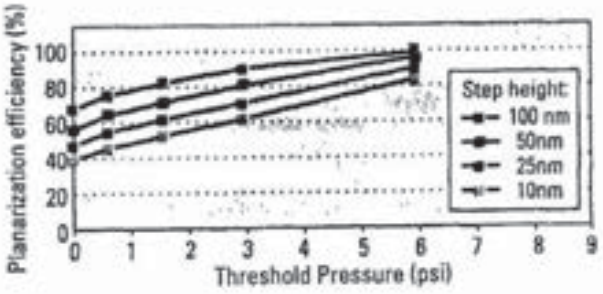


Fig. 8-20 Increasing the threshold pressure improves planarization efficiency for various final step heights (500-nm initial step height).[16]

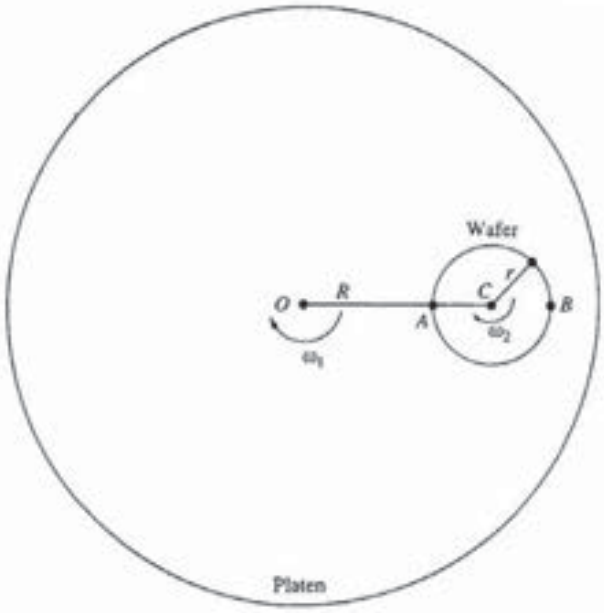


Fig. 8-21 Plan view of a single-carrier single-platen rotary-polisher, showing relative rotary motions of both the carrier and the platen.

Now assume that the platen and the wafer carrier both rotate. In this case (referring to Fig. 8-21), it can be noted that the centers-of-rotation of the wafer-carrier and the platen are offset from one another. If the wafer-carrier and the platen have the same rotational frequency, and are rotating in the same direction, point B of the outside edge of the wafer in that figure experiences zero relative velocity, and the point A on the inside edge of the wafer experiences the greatest relative speed of the surface of the wafer to the pad. Thus, the instantaneous removal rate at the edge of the wafer is greatest when the edge is directly between the two centers (point A), and zero when in the extreme position (point B). However, since all points on the edge of the wafer pass through points A and B at a constant frequency, the time-average RR at the edges will be the same as at the center of the wafer. Thus, it is the average-radius from the center of the platen, r_{av}, that defines the distance the wafer travels and the average speed of travel, S_t, for purposes of the total film removed and RR calculations:

$$S_t = 2\pi r_{av} f\,(\text{rpm}) \qquad \textbf{(8.18)}$$

In practice, within-wafer-non-uniformity of RR is observed in rotary-CMP tools (e.g., WIWNU). One form of such WIWNU is a spoke-like pattern due to vibration harmonics. Driving the carrier in rotational synchronization with the platen usually causes the worst vibrations. Hence, the wafer-carrier is usually driven at a constant difference in frequency to the platen. Since the correct wafer-carrier rotational-frequency for perfect circular symmetry is not known, it must be experimentally determined.

8.5.2 Models Which Describe Factors that Impact Oxide-RR in CMP Through the Preston Constant

There are many factors that impact the oxide-RR in CMP besides the applied pressure and velocity. These are expressed in Preston's Law by changes in the Preston constant. Here two models are given that have been proposed to explain the dependence of oxide-RR on other phenomena besides pressure and velocity.

8.5.2.1 Impact on the Oxide-RR from the Electrochemical Interactions Between the Slurry Particles and the Oxide Surface: A model was proposed by Cook in which he postulated that a mechanism for oxide-removal by CMP involved contact-mediated reactions with the solvent (water).[17] That is, local events in CMP may be considered to be the approach to contact of two opposing asperities: **1**) pad plus adhering particle; and **2**) the oxide surface. Asperity-contact causes reactions in the vicinity of the contact, and these result in mass removal from the oxide. The depth and width of the swath removed is determined by the size of the slurry particle and the chemical-reaction-rate, which is largely controlled by the interaction force between the contacting surfaces. The length of the swath gives an indication of the duration of the asperity contact.

For the case where the slurry particle and the surface are of the same material (as occurs in oxide-CMP with a silica-based slurry), the forces due the Zeta-potential act to repel the silica particles from the oxide surface, while the Van-der-Waals forces cause the oxide surface to attract the slurry particles (see Sect. 8.7.1 for a discussion of the Zeta-potential and the Van-der-Waals forces). The curve that graphs the force-versus-distance for such a case of opposing forces between the particles and the wafer-surface exhibits a maximum repulsion at some characteristic separation distance. This may be perceived as an activation-energy-barrier for closest approach. The particle must thus have a kinetic-energy greater than this activation-energy-barrier in order for direct contact to occur between the slurry particle and the oxide-surface. A reduction in this activation-energy-barrier should result in a higher probability of contact, and therefore cause an increase in oxide-RR.

Based on this model, Cook suggested that an increase in RR could be achieved by using a slurry whose particles were different than that of the oxide (e.g., use CeO_2 slurry particles instead of silica), and in which the Zeta-potential had a value that caused the slurry particle and the oxide to attract one another. This would strongly increase the probability of contact between the slurry particle and the oxide-surface and thus increase RR.

8.5.2.2 Dependence of Oxide-RR in CMP on Dielectric Hardness: Dielectric materials vary in hardness from the very-hard and inelastic silicon-nitride, to softer dielectrics such as doped-oxides and spin-on-dielectrics (SODs). The RR of dielectrics has been found to increase with a decrease in their hardness. Figure 8-22 shows that as the hardness of fluorinated-silicate-glass (FSG) is decreased, the RR increases.[18] (The hardness of the FSG in this study was decreased by increasing the concentration of fluorine that is incorporated in the oxide.) In another work, the RR of PSG and BPSG was studied as a function of the P-concentration in a P-doped oxide-film, and as a function of the B- and P-doping in a BPSG oxide film (Fig. 8-23).[19] Higher doping is seen to cause the RR to increase. This is correlated with oxide hardness, since the incorporation of P_2O_3 and B_2O_5 into an SiO_2-network causes a softening of these oxide-films.

8.5.3 Models Which Describe Chemical-Mechanical Phenomena that Give Rise to Planarization Phenomena Associated With CMP of Oxides

Various models have been proposed to explain how chemical phenomena act in concert with the mechanical forces affecting oxide-RR (as expressed by Preston's Law). These models also provide insight as to why oxide-CMP can selectively remove the "up-regions" of an oxide film on wafers faster than the "down-

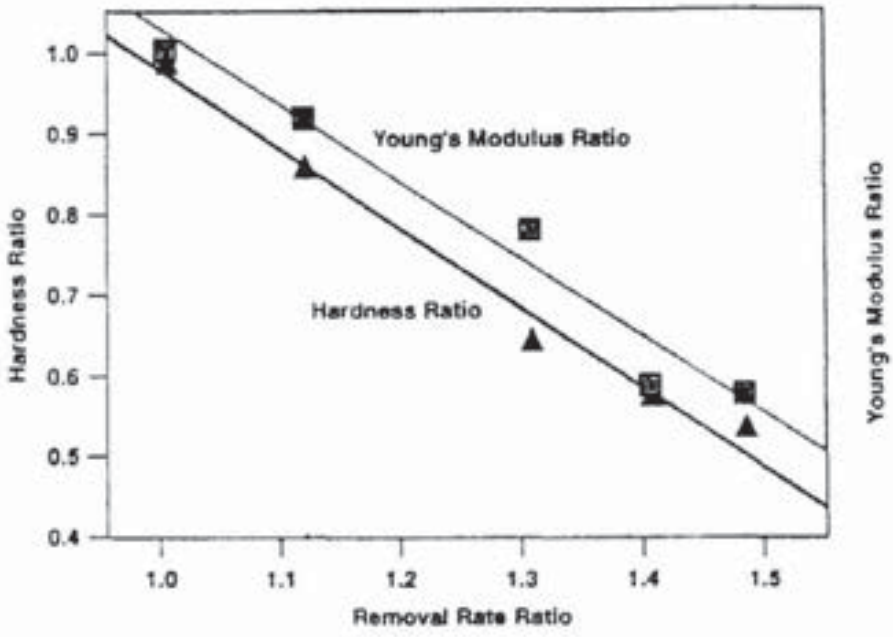


Fig. 8-22 A linear relationship exists between the CMP-removal-rate and the hardness and elastic modulus of F-SiO_2. Removal-rate, hardness, and elastic modulus are all normalized with respect to those parameters of undoped TEOS-oxides.[18]

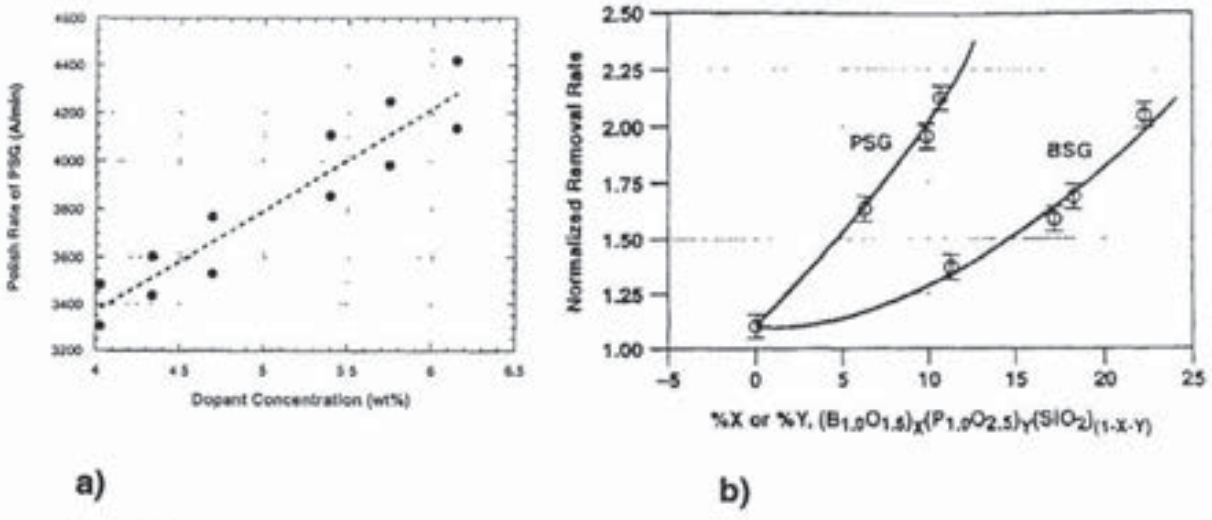


Fig. 8-23 a) Removal-rate of PSG as a function of phosphorus-concentration: pressure 4-psi, table speed: 40-rpm, carrier speed: 643-rpm; b) Removal rates of P-doped oxides and B-doped oxides as a function of P- and B-doping-concentration in the oxide.[19]

regions" (thereby allowing planarization of the film to occur). Some ideas from these models are discussed in this section.

Izumitani proposed that optical-glass-polishing "proceeds by the formation of a hydrated-layer by means of a chemical-reaction between the glass-surface and water and then the removal of this hydrated-layer by the abrasive particles."[20] This proposed mechanism is consistent with the observation that the polishing-rate is extremely low without the presence of water to cause the hydration. (Polishing with a slurry of silica-particles in a non-aqueous solution [e.g., in oil] exhibits substantially lower polishing rates.) The hydrated-layer that thus forms on the oxide-surface generally has a much lower hardness and mechanical strength than the SiO_2 beneath it. Therefore, plastic-deformation results in mechanical removal of the hydrated-layer (by a plowing effect) when it is rubbed by the abrasive particles.

The same type of mechanism is expected for oxide-CMP. Removal of the oxide takes place by a combination of mechanical and chemical action, using an abrasive slurry dispersed in alkaline solution (Fig. 8-24). Since silica-particles are used in the slurry of oxide CMP, they have about the same hardness as the SiO_2. Thus, the unreacted SiO_2 is not appreciably removed (nor damaged) by the plowing effect when it is rubbed by the silica particles in the slurry. However, the silica-slurry-particles still cause the oxide to plastically deform. This deformation

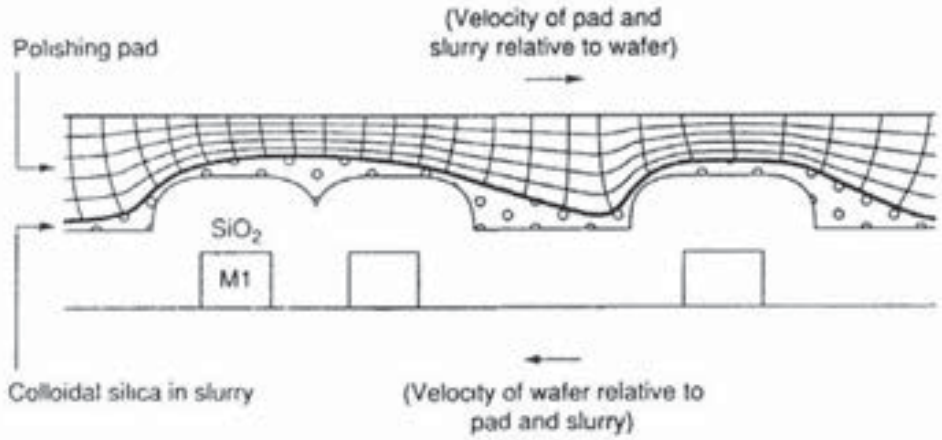


Fig. 8-24 Depiction of polishing-pad compression and localized pressures applied to topographic features. Pressures are greatest at the top-leading-edges of steps, and are lowest in the wakes of the steps.

is likely to be assisted by frictional heating. It has been determined that oxide readily undergoes hydration (water entry) if it is plastically deformed by a diamond indenter. When entered by water, the hardness and mechanical strength of the oxide is generally decreased. Thus, during CMP, it seems that hydration of the oxide film surface is caused by the plastic deformation of the oxide by the abrasive particles. The degree of deformation is increased as the pressure of these particles on the surface is increased, and by frictional heating. The softer hydrated-layer is then removed by the plowing action of the abrasive particles, which increases as the linear-velocity of the polishing pad is increased. These conclusions are substantiated by Preston's Law. The properties of the pad also appear to play an important role in CMP. It is probable that the polishing pad has to be soft enough to grip the abrasive particles, but hard enough to transmit the pressure to the oxide. It is generally agreed that some sort of chemical mechanism (such as SiO_2-dissolution) is a cofactor in causing oxide-CMP. Various chemical processes have been proposed.[21] The larger polishing rate in higher-pH slurry is often cited as evidence for silica-dissolution.

8.5.3.1 Model of the Mechanism that Produces Planarization of Steps in Oxide CMP: While the above model gives some insight into how oxide is polished with a silica-based slurry, it does not address the issue of how planarization of the oxide by such polishing is achieved. To obtain planarization of a film with topography, the surfaces of elevated features ("up-areas") must be eroded faster than those at lower elevations ("down-areas"). The selective removal of material in "up-areas" is empirically observed to depend on many factors, including the

feature dimensions, the structure (and surface condition) of the polishing pad, and the height of the step being removed

It is important to note that the case being discussed here involves the CMP of an oxide layer that proceeds only until the oxide surface is planarized, and that the surface of the wafer after the CMP step is *homogeneous* (i.e., only oxide is exposed over the entire wafer surface at the conclusion of the CMP step, as shown in Fig. 8-25a). This differs from the case when a film is polished until it is removed from the surface in some regions, but is allowed to remain in recessed regions of the wafer surface after the CMP step. In that case (which is described in a later paragraph), the wafer surface is *heterogeneous*, insofar as some exposed areas of the wafer contain the polished film residing in recesses, and others contain material that underlaid the film that was polished (Fig. 8-25b). Note that in the second case, the CMP process is essentially a *blanket-removal-process of a planar film* because the film being removed is already roughly planar when the CMP process is started.

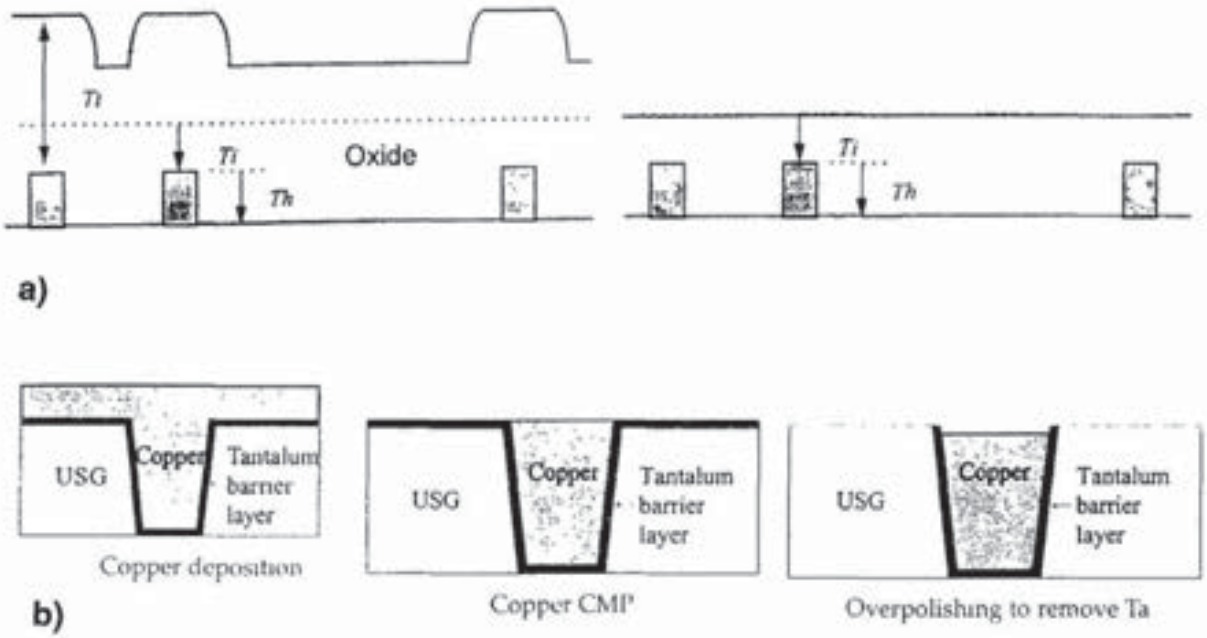


Fig. 8-25 The process of removing material by CMP falls into two categories: a) Removing homogeneous material and stopping while the removal of the surface film is incomplete (e.g., planarization of an ILD oxide); b) Removing deposited material, but continuing until all material is entirely removed on some regions of the wafer surface, while in other regions it remains recessed below the surface (e.g., Cu-damascene feature formation by CMP).

Selective removal of material that protrudes above the surrounding topography (as shown for the case of Fig. 8-25a) is thought to arise in part from the compliant nature of the polishing pads.[22] Figure 8-24 depicts the localized compression occurring within a pad as features on the wafer surface are driven across it. The pad applies pressure to the wafer surface through the slurry, with the largest pressures being exerted where the pad is most compressed. Higher pressure is applied to elevated regions ("up-areas" - Fig. 8-26) and the leading edges of the features that are moving against the compressed pad, compared to lower-lying regions ("down-areas") and trailing edges.[23] The difference in local applied pressure causes the material removal rate to be fastest at the upper corner of the leading edge of a step. Figure 8-27 shows how a feature would be planarized as the leading edge is eroded away as a function of time during a CMP process. (In the next section it will demonstrated how this phenomenon - in concert with wafer rotation during polishing - causes smaller features to be rounded-off and polished-aster than wider features.)

As the steps erode, the "down-areas" are exposed to increasing pressure, and the removal-rate there increases (approaching that of the "up-areas"). As this happens, the rate of step-height-reduction decreases toward zero. In the latter stages of the polishing process the selectivity is so low that small residual steps remain, and in practice cannot be fully removed by continued polishing.

In the second case described above (Fig. 8-25b), the material of the blanket-deposited layer is removed non-selectively by CMP. This non-selective removal continues until all of the film is polished away from the areas of the substrate, and only film material remains in the trenches. However, if the CMP process is continued beyond that point, material within the trenches will continue to be removed. This removal will be greater in wide trench areas than in narrower trenches. The result of such removal from wide trenches gives rise to a CMP problem called *dishing*. The phenomenon of dishing and methods developed to mitigate it are also discussed in other parts of the book. In Sect. 8.12.1, the general aspects of dishing are covered, and the impact of dishing effects on the fabrication shallow-trench-isolation structures is found in Chap. 9. Chapters 15 and 16 provide additional information on the impact of dishing on Cu-CMP and on the CMP processes used to polish inlaid metal lines in damascene processes.

8.5.3.2 Pattern-Dependence of the Planarization Rate in CMP Processes: The degree of planarization (DOP) of features for the CMP processes depicted by Fig. 8-17a has been found to depend on the feature sizes and the density of features within a region of a die. This DOP dependency on feature size and pattern density is one cause of WIDNU. The reasons for such WIDNU are described here.

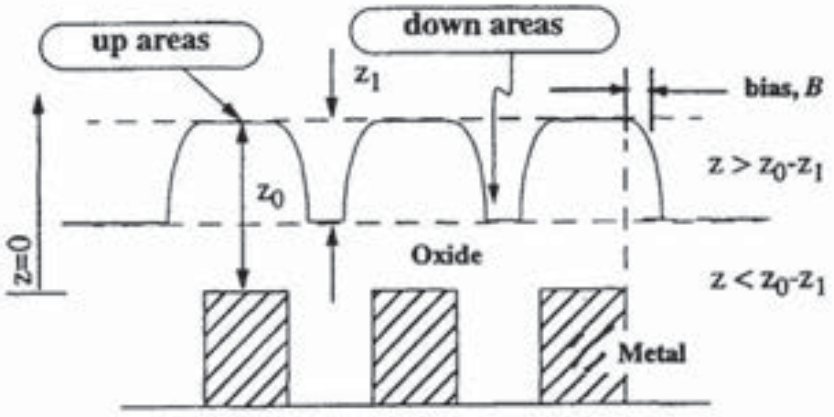


Fig. 8-26 Cross section of conformal dielectric on patterned metal, showing the "up-regions" and "down-regions" of the dielectric film.

The first question to answer is why narrow, isolated-features tend to planarize faster than wide isolated features. From Fig. 8-27 we see that the leading edge of a step will be removed as a function of time in an oxide CMP process. However, in a process in which the wafer carrier is rotating, all top edges of the feature being polished become the "leading edge" as the feature is rotated by the carrier. As a result, the entire feature is planarized according to the sequence shown in Fig. 8-28a. If the feature is wide, the leading edges will be polished away, as shown in Fig. 8-28b, but the material in the center of the wide feature will not be polished as quickly as material at the feature edges. As a result, wider, isolated-features will planarize more slowly than narrow, isolated features. This will cause such WIDNU as shown in Fig. 8-13.



Fig. 8-27 A depiction of how the leading edge of a feature of an oxide step (i.e., the edge that is pressed against the pad by the combined movements of the carrier and platen), is eroded by CMP.

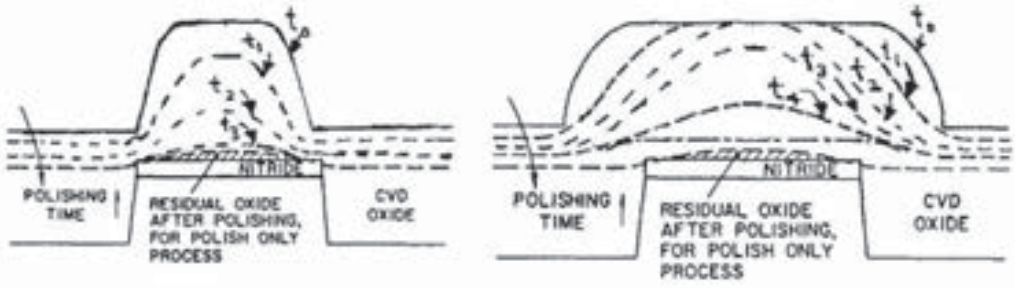


Fig. 8-28 Schematic cross-section of a narrow and wide feature of oxide being polished by CMP, as a function of time. Since the leading edges are polished most rapidly (and the feature continually rotates, so that all edges become the leading edge for the same amount of time), the narrow feature is planarized more rapidly than the wide feature.

The next question is: Why are densely-spaced narrow features planarized more slowly than isolated, narrow features? As first discussed by Burke,[24] the time required to planarize an "up-area" structure as shown in Fig. 8-26 (or to reduce its height to some fraction of its initial value) is dependent on the "down-area" adjacent to it. Specifically, the time to planarize a step depends on the width of the "up-area," the width of the "down-areas" between "up-areas," and the ability of the polishing-pad to rigidly span that distance between "up-areas." The stiffer the pad, the more rigid the span - and the slower the rate of material removal from "down-areas." This gives good planarization. The softer the pad, the less able it is to bridge larger distances. Thus, more material is removed by softer pads in "down-areas" between adjacent "up-areas." This gives poorer planarization. When features to be polished are separated by distances greater than the leveling-distance of the pad, there is no selectivity between "up-areas" and "down-areas," and planarization does not occur. In practice, some selectivity in RR occurs between "up-" and "down-areas," but this selectivity is not complete. Ideally, the leveling-distance of a pad should be large enough to planarize the widest "down-areas" on a chip. On the other hand, the greater the leveling-distance, the more a pad will respond to wafer-scale non-flatness parameters, such as TLV and bow. This results in non-uniformity of the polished film thickness across the wafer (a subject discussed in more detail in Sect. 8.12.2).

In summary, planarization will occur most slowly for step heights in an oxide over the largest or densest interconnect structures adjacent to wide unpatterned

field areas, scribe lines, and wafer edges. It will occur most rapidly for step heights in the oxide over smaller arrays and narrow, isolated features. It is also necessary to specify an acceptable degree of planarization, because in the latter stages of the CMP process, material is removed nearly as fast from "down-areas" as from "up areas." Step heights no longer decrease at appreciable rates.

Like metal-CMP processes, oxide-CMP in a production environment is often a multi-step process. When in-line endpoint-detection is employed (as will be discussed in Sect. 8.6.9), two or more polishing steps may be needed to obtain the correct oxide thickness. In addition, a separate final-polish (or buff) procedure is frequently used to provide light polishing and to remove slurry particles and small wafer-defects. A small amount of oxide can also be buffed away with this buffing procedure (typically 30–40-nm), which aids in removing defects (especially residual slurry particles and the small scratches created by the primary polish). Use of a buff procedure is one of the reasons why CMP can actually *reduce* defect-densities (including eliminating some types of defects created in other processes), rather than increase them.[25]

8.5.4 Metal-CMP Mechanisms

The mechanisms involved in the CMP of metals which cause planarization of the metal film surface, follow a relatively simple three-step sequence proposed by Kaufman.[26] This three-step sequence is as follows (Fig. 8-29):

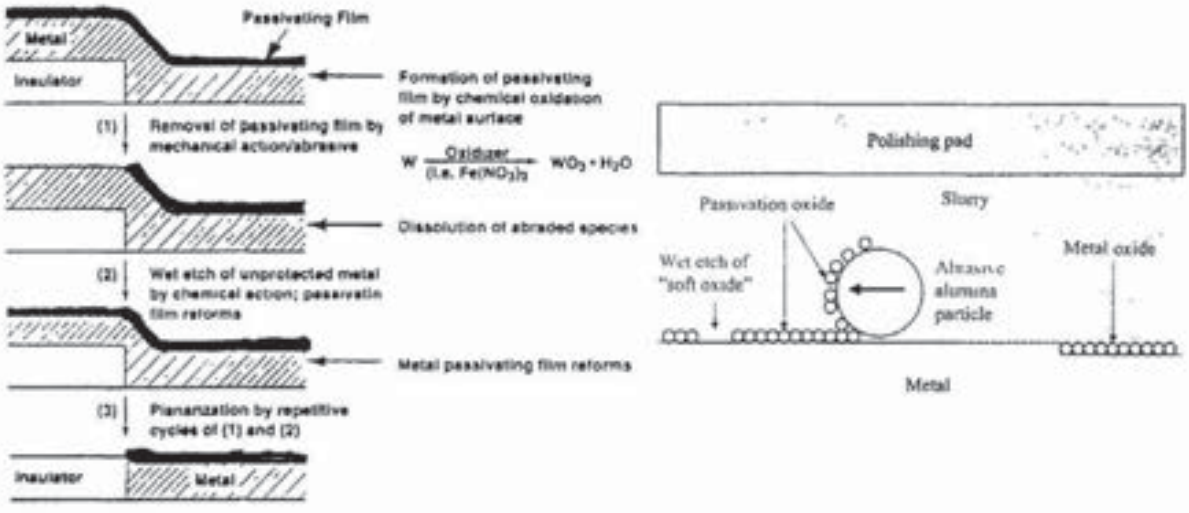


Fig. 8-29 Schematic representation of the model of CMP of metal films.[26]

1. A passivation layer (e.g., an oxide of the metal) is chemically formed on both the "up-areas" and "down-areas" of the metal film, due to the oxidation potential of the solution. (For effective polishing, it must be possible to both oxidize the metal and closely match the hardness of the slurry abrasive to the hardness of the material being removed.)

2. This freshly formed oxide film is softer than the metal itself and is mechanically removed from the "up-areas" by the pad/abrasive. (While the "down-areas" are not contacted due to the partial flexibility of the pad.)

3. The corrosive nature of the slurry re-grows the passive film on the "up-ares," while the corrosive elements are blocked from the metal in the "down-areas" by the passivation film.

The "up-areas" are continually corroded and polished until the pad flexibility allows it to contact the "down-areas." At that point, the planarization process again depends on the hardness of the pad and the down-force applied to the wafer. Thus, the corrosion properties of the slurry are chosen to convert the metal surface into a film that is soft enough to be abraded by the pad-plus-abrasive-material, yet hard and impermeable enough to prevent outright etching of the metal.

The goal is to find slurries with just the right combination of chemical and abrasive aggressiveness to produce high removal-rates, good-planarity, local-uniformity, and high-selectivity. Although the metal-CMP-model is simple to understand, successfully carrying out the above process for specific metals can be challenging. The slurry's chemical-composition tends to play a more important role in metal-CMP than in dielectric-CMP. As a result, it is possible to achieve fairly high metal-to-oxide selectivity (e.g., 120:1 for tungsten to silicon dioxide). However, the slurry solution may also etch the metal-oxide that is formed. A careful balance between etching and oxidation must therefore be maintained. If etching dominates, then the metal will etch isotropically and planarization will not be obtained. The metal-deposition-step may also have to be modified to prevent metal deposition near the wafer edge. Otherwise, metal residue may re-main on the wafer's beveled edge after CMP, causing yield and reliability problems.

Pourbaix diagrams are often used to identify which pH-range is most ideal for CMP of metals.[27] *Pourbaix diagrams* are graphical representations of the domains of stability for different oxidation states of metals in aqueous solutions with respect to pH and electrochemical potential (i.e., contour maps in a solution potential versus pH). For example, the Pourbaix diagram for W (shown in Fig. 8-30) indicates that in order to form an un-dissolved protective WO_3 layer over the W, the pH of the solution should be below 4, and have a slightly positive oxida-

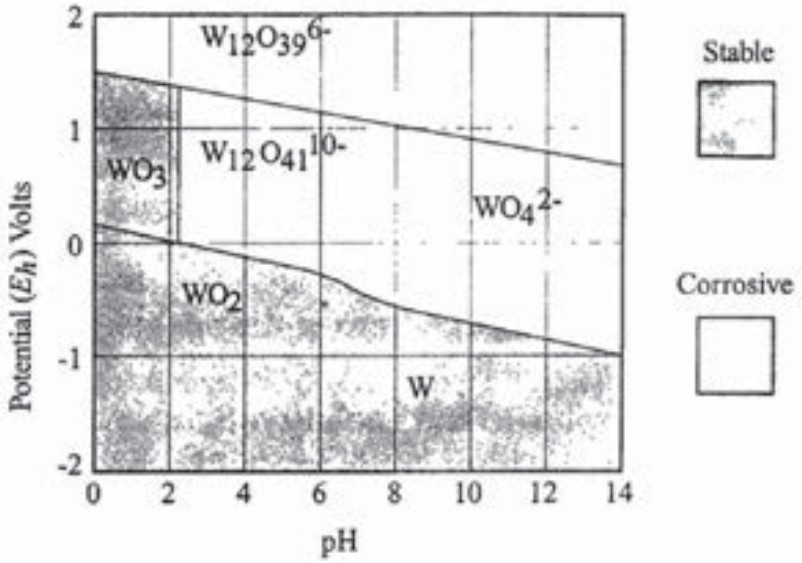


Fig. 8-30 Tungsten solubility versus solution-pH (Pourbaix) diagram.[27]

tion potential. For this reason, most tungsten slurries have a low pH, typically with pH-values between 2 and 3.

Unlike oxide CMP processes (whose purpose is to remove a specific amount of the dielectric layer), the metal-CMP processes require some degree of over-polish to ensure that metal is completely removed from field areas. The challenge here is to optimize such overpolish so that process time and oxide loss are minimized (i.e., to maximize throughput), while compensating for any non-uniformities from prior steps (and without causing any new uniformity problems from dishing or erosion).

It should also be mentioned that one aspect which makes metal-CMP processes easier to control than oxide-CMP processes, is that commercially available endpoint methods are much more sensitive for metal CMP. That is, metal-CMP processes are carried out by polishing-back blanket-metal-layers until all of the metal on the wafer surface is removed (but the metal in the via holes or within the trenches remains). Once the metal on the top surface is removed, the underlying oxide in the field is exposed. Consequently, this endpoint is easier to detect by monitoring a variety of phenomena associated with the clearing of the metal over the oxide, such as: **a**) a change in surface reflectivity; **b**) a change in surface temperature; or **c**) a change in the torque of the platens or wafer carrier (due to the different frictional coefficients of the materials being polished). In any case, the optimum metal-CMP process would have a clear endpoint-signal -followed by a short overpolish step (to remove any

metal remaining in dished areas of the wafer, or in surface scratches). The result should be a surface in which the metal and the barrier-layer is completely removed, with a minimum of the underlying dielectric being eroded, and where a minimum amount of dishing is produced.

8.6 CMP EQUIPMENT

There are number of different tools and consumables used in a complete CMP process, including:

1. CMP polishing tools;
2. Slurries;
3. Pads;
4. Slurry distribution systems;
5. Endpoint detection systems;
6. Cleaning tools;
7. Metrology tools.

These items are described in the following sections.

8.6.1 CMP Polishing Tools

The basic polishing tool used for CMP evolved from the machines used to polish bare-Si-substrate-surfaces. When the idea for CMP was first conceived at IBM, they purchased silicon polishing tools (used by wafer manufacturers) from Westech (now SpeedFam-IPEC) and Strasbaugh. The basic configuration of the wafer-polisher was retained, but the tools were modified to make them suitable for developing CMP. As mentioned earlier, in such polishers the wafer is mounted upside down on a wafer-carrier, and the carrier is pressed downward against a polishing pad (Figs. 8-31 and 8-32). Both the wafer carrier and polishing pad are rotated. Slurry is dripped onto the polishing pad, and it becomes saturated with slurry. The pad thus drags slurry beneath the rotating wafer carrier, where it can polish the wafer surface (as described earlier in the section on mechanisms of CMP). The first polishers adapted for CMP had a single wafer-carrier.

The CMP-polisher differs from the wafer-polishing tool in a number of ways. In polishing silicon-wafers, several tens-of-microns of material are removed, while the CMP-process removes only about 0.5–1.0-μm of material. The uniformity of material removal across a wafer in CMP is also much more stringent than in silicon-wafer polishing (±0.5–0.1-nm for CMP). Furthermore, CMP-tools require a higher-level of automation, throughput, reliability, and critical process-parameter-control. Thus, despite more than 30 years of experience in silicon-wafer polishing, new processes and equipment had to be developed by the IC industry for carrying out CMP.

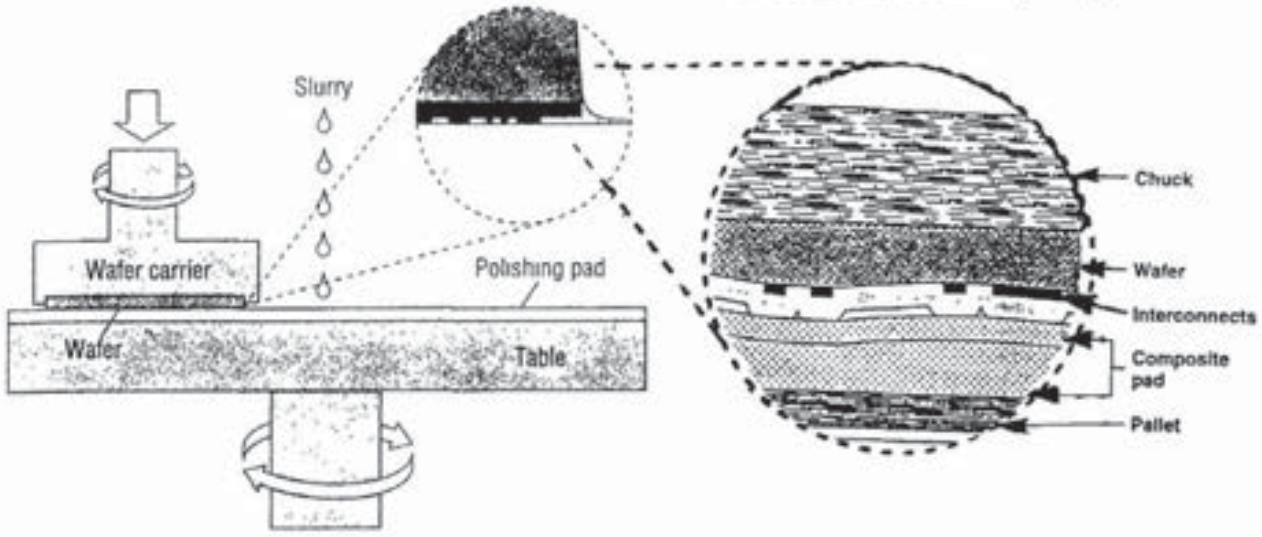


Fig. 8-31 Schematic diagram of a wafer being polished by CMP.

The Westech and Strasbaugh polishers first used by IBM were also the tools that found widest use in the early CMP processes. IBM provided to SEMATECH one of these silicon-wafer polishing tools from Westech. This tool was used by SEMATECH to launch a program that focused on improving the tool's ability to do CMP on the thin films encountered in VLSI structures, and to increase the tool reliability. The results of this program were used in designing the IPEC model 472®-CMP-tool, which achieved wide adoption for CMP applications during the mid-1990's. Figure 8-33 shows a plan view of an IPEC Avanti 776® polisher, a modified version of the 472 tool.

8.6.1.1 CMP Polishing-Tool Configurations: While early CMP tools used a single wafer-carrier, more recent designs contain multiple-carriers (and multiple-platens). This increases the throughput of the polisher and allows multi-step CMP processes to be carried out on a single tool. The classes of CMP polishing tools can be categorized according to the manner in which motion between the wafer-carrier and polishing table is produced. These categories are as follows:

1. *Rotary CMP Tools* (in which both the table [or platen] and the carrier rotate);
2. *Orbital CMP Tools* (in which the carrier rotates, but the table moves in an orbital path);
3. *Linear CMP Tools* (in which the carrier rotates and the table moves in linear path);
4. *Fixed Table CMP tools* (in which the table is stationary, but the carrier moves in an orbital path).

Table 8-1 lists some commercial CMP polishing tools and categories into which they fall.

8.6.1.2 Rotary CMP Tools: Figure 8-32a shows a schematic of a single-carrier/single-platen *rotary CMP tool.* The motion between the wafers and the pad is created by the rotation of the platen together with the rotation of the wafer-carrier. (The carrier also holds the wafer face down, and applies a downward force against the surface of the pad mounted on the rotating platen.) The pad is afixed to the flat table by a waterproof-adhesive. The following tools are examples of rotary-motion CMP tools with multiple wafer-carriers: Strasbaugh Symphony,® which has four wafer-carriers and three platens (Fig. 8-34); SpeedFam-IPEC Auriga® which has five wafer-carriers and two platens (for primary and secondary polishing, Fig. 8-35); Cybeq Nano-Tech-Isoplanar-8000® which has six-wafer carriers; and Applied Materials Mirra® with four wafer-carriers and three platens (each with an *in-situ* pad-conditioning arm, Fig. 8-36). In the Mirra,® a central spindle transports the wafer-carriers from one platen to the next. A three-platen-configuration (such as is available on the Symphony® and the Mirra®) permits three-step CMP-processes to be performed on a single tool.

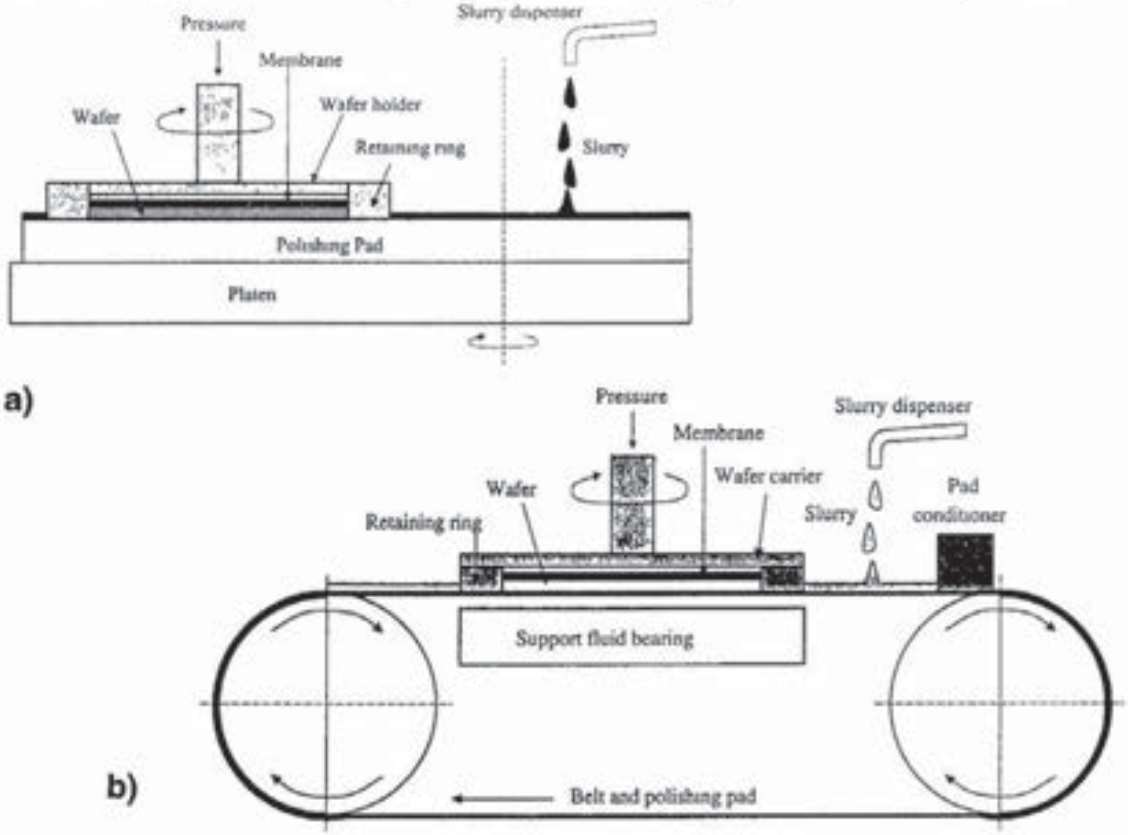


Fig. 8-32 CMP tool configurations: a) Rotary tool; b) Linear tool.

TABLE 8-1 Classification of Some CMP Polishers

Vendor & Model	Table motion–number	Carrier type-number	Built-in endpoint methods1	Special features
	Established	**Polishers**		
AMAT Mirra	Rotary–3	BB-CRRC–4	Red laser reflectometer	Sequential use of 3 tables
Cybeq 8000	Rotary–1	BB-CRRC–6		
IPEC 372/472	Rotary–2	FBC/NRRC–1	Absolute temperature	Sequential use of 2 tables
IPEC 676	Orbital–4	BB-CRRC–4		Parallel tool
IPEC 776	Orbital–4	BB-CRRC–4		Ditto, Integrated Ontrak scrubber
Speedfam/ Auriga	Rotary–1	FB-NRRC-5	Visible light reflectometer	
Strasbaugh/6DS-SP	Rotary–1 or 2	FB-NRRC-2		2 carrier size buffing platens, optional 2nd table
Strasbaugh/ Symphony	Rotary–3	FB-CRRC–4		
	Developing	**Polishers**		
Lam/Teres-Ontrak/Synergy	Linear-2 Integrated Cleaner	?–2		Fluid bearing platen
Obsidian	Fixed–1	Orbital, BBC?–1		Disposable 3-M slurryless pad

8.6.1.3 Orbital CMP Tools: In *orbital CMP tools*, the wafer-carrier rotates in the same manner as in the rotary CMP tools. The platen, however, executes an orbital path about the center of the rotating carrier (instead of rotating about its own central axis (Fig. 8-37). In this approach the removal-rate-non-uniformity problem encountered in rotary CMP tools is circumvented. This occurs because the travel distance of all loci on the wafer surface relative to loci on the platen surface are equivalent. Another advantage of this tool configuration is that it uses a platen with a smaller radius than that needed in a rotary polisher. As a result, the footprint of the polisher can be significantly reduced. Orbital CMP tools are sold by IPEC-Speedfam in their AvantGaard 676,® AvantGaard776,® and Momentum® models. Figure 8-38 shows the SpeedFam-IPEC Momentum,® which has four wafer carriers and four platens.

8.6.1.4 Linear CMP Tools: CMP tools that have a pad in the form of a continuous belt (that moves in the same fashion as a treadmill) are also available. The pad is thus moved past the rotating wafer carrier in a linear (rather than rotary) motion (*linear CMP tools*). These tools offer the potential benefits of higher linear

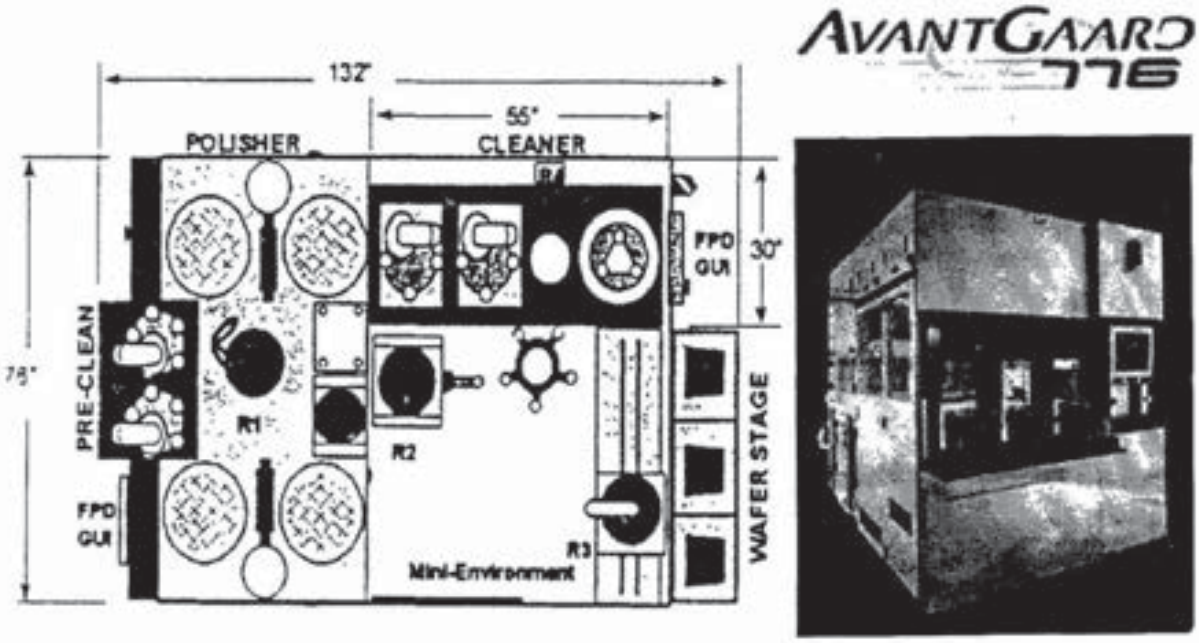


Fig. 8-33 Plan view of an IPEC *Avanti* 776® CMP tool. Courtesy of Speedfam/IPEC.

linear speeds 400–500-ft/min (versus 34-ft/min in rotary CMP tools), harder pad materials, and reduced polishing-pressure (i.e., because the polish rate can be maintained at a high value with reduced applied pressure, as a result of increased pad velocity). This type of tool may be able to produce better planarization of oxides in a shorter time and reduce the problem of dishing in CMP (see Sect. 8.12). Linear CMP tools are offered by Lam Research (Teres® Fig. 8-39), Aplex, and Doosan. The Teres® has two linear polishing belts, and one station with a rotating platen (for carrying out buffing or cleaning steps). The wafer carriers move independently from one station to another. The Teres® also employs an *air bearing* beneath the polishing belt. This allows the pressure distribution across the wafers to be profiled. At a given pressure and speed, adjustments to the pressure-distribution profile can be made to significantly reduce film-thickness non-uniformities (Sect. 8.12). The removal-rate non-uniformity problem encountered in rotary CMP tools is also circumvented in linear CMP tools. Improved planarization efficiencies (i.e., less oxide needs to be deposited to achieve complete planarity) have also been claimed for this configuration.

8.6.1.5 Fixed-Table CMP Tools: A CMP tool that employs a stationary platen and a wafer-carrier that moves in an orbital path over the platen (Fig. 8-40) was announced by Obsidian Corp. in 1999. (Another possibility would be to rotate the carrier while at the same time moving it linearly back and forth across the table,

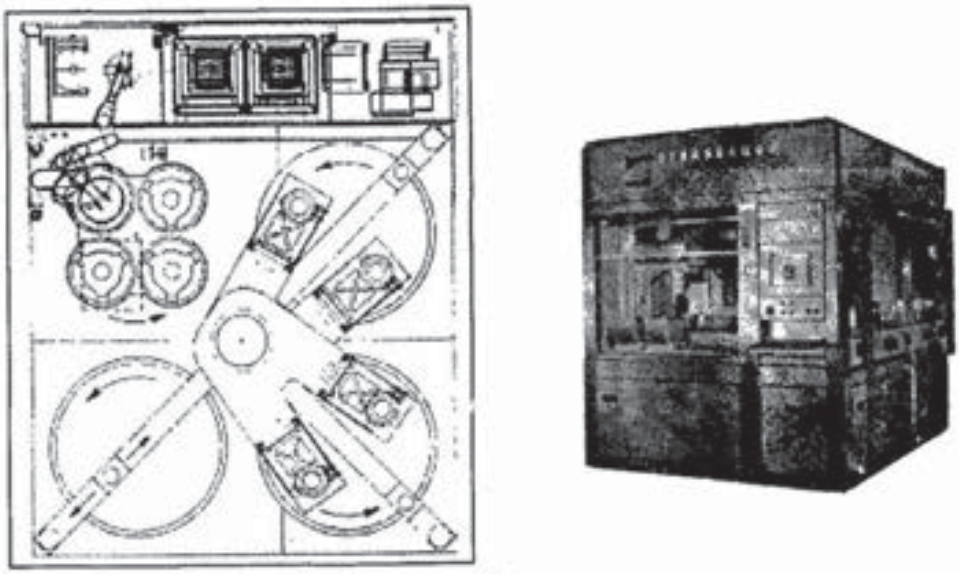


Fig. 8-34 Schematic drawing of the Symphony® CMP tool. Courtesy of Strasbaugh Corp.

see Fig. 8-41). This tool also employs a slurry-less pad supplied by the 3M Corp. The pad material is mounted on a roller (akin to a roll of paper towels). After each wafer is polished, the roll is advanced to remove the used pad-surface. This makes a new segment of pad material available to polish the next wafer (roll-to-roll transport of the slurry-free pad). Applied Materials acquired Obsidian in 2000, but by the end of 2000, the tool had not yet been commercially released.

8.6.1.6 300-mm CMP Tools: Many semiconductor equipment manufacturers are readying CMP polishing tools for 300-mm wafers. By mid-2000 more than a dozen CMP equipment manufacturers have exhibited or announced intentions to participate in the 300-mm market.[27] It is reported that most 300-mm CMP tools are capable of reaching the performance capabilities of the 200-mm tools. However, challenges remain with respect to the issues of consumables costs, integration of metrology, and full automation. In some cases a reduction in consumables cost is being achieved. For example, in the Speedfam-IPEC 300-mm polishing tool, slurry use is increased by only 20% over that needed in the 200-mm tool, whereas the area of the wafer being polished is increased by a factor of 2.25. This has a net positive effect on the CoO. Edge exclusion zones of 3-mm or less are also reportedly being achieved on these tools.

8.6.2 Wafer Carriers in CMP Polishing Tools (and the CMP Edge Effect)

In Fig. 8-42 the wafer is shown mounted on the carrier. The traditional wafer carrier consists of a circular plate (chuck) within a pocket (which holds the

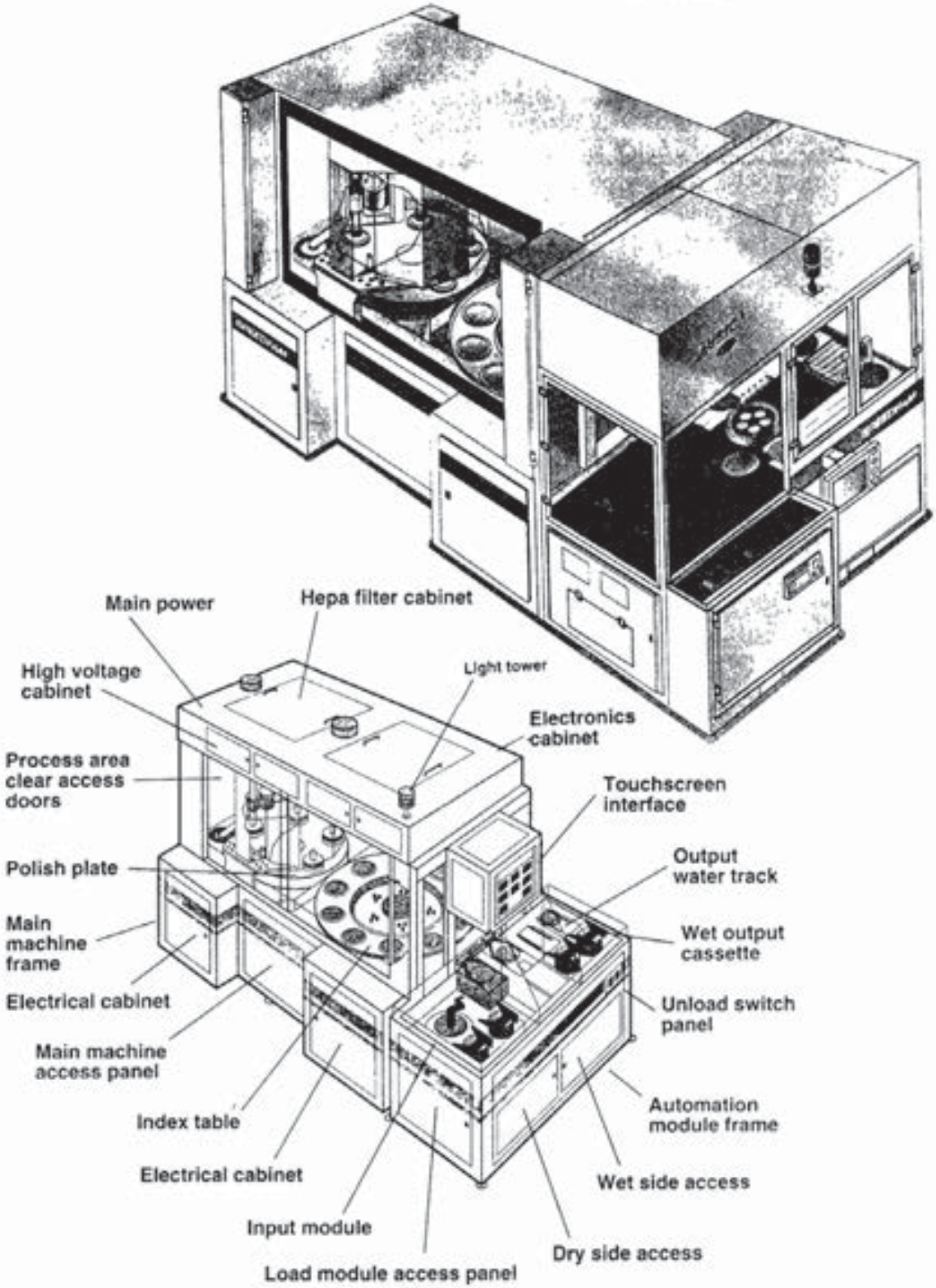


Fig. 8-35 Detailed view of the SpeedFam-IPEC Auriga® five-carrier polishing tool. Courtesy of SpeedFam-IPEC, Inc.

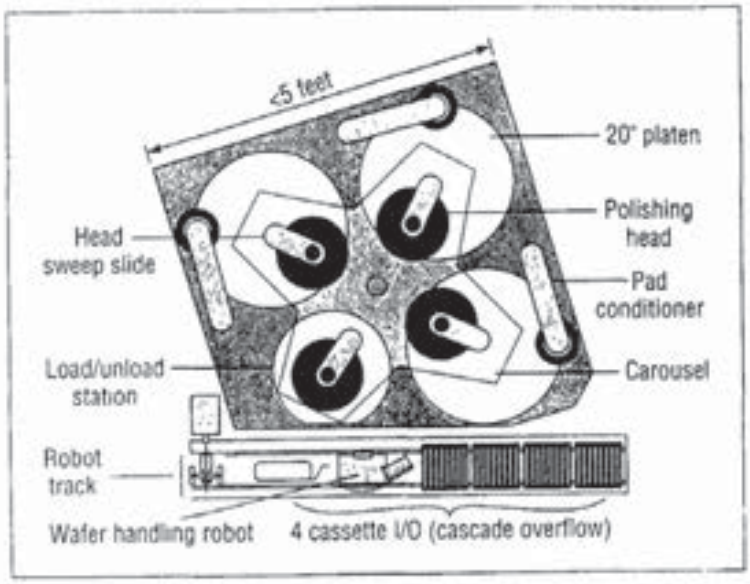


Fig. 8-36 Architecture of the Mirra® three-platen CMP tool. Courtesy of Applied Materials, Inc.

wafer), attached to a rod by a gimbal (Fig. 5-43)). Downward force is applied to the plate by the rod during polishing. The depth of the pocket in which the wafer resides is chosen to allow the top-surface of the wafer to extend beyond the pocket, permitting its surface to be polished. Instead of the wafer being directly in contact with the steel chuck, a backing film is sandwiched between the wafer and the chuck. This film provides elasticity. Without a flexible backing film, any

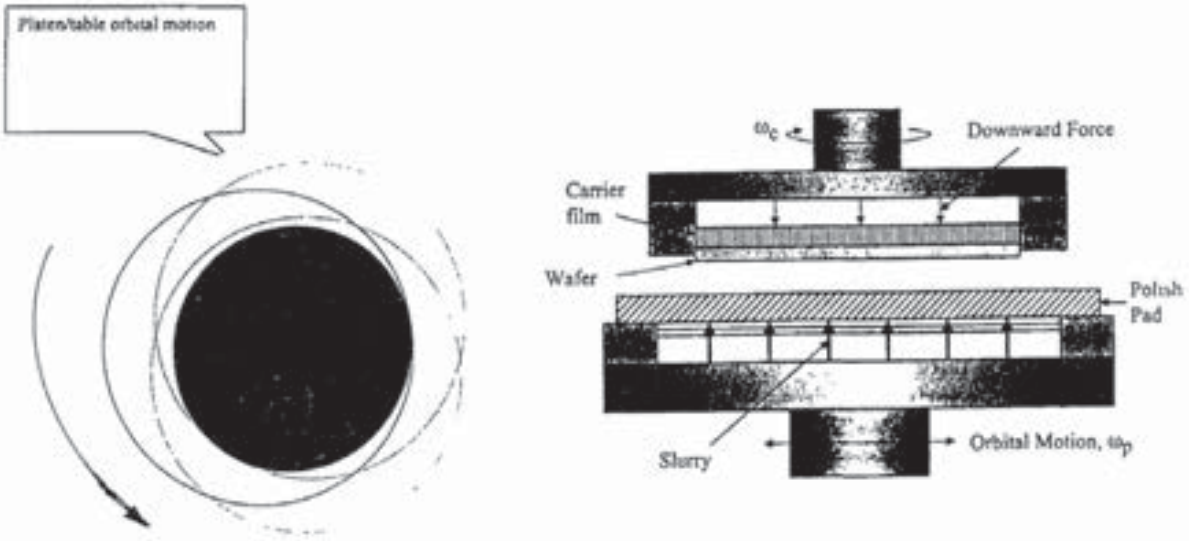


Fig. 8-37 Schematic drawing of orbital platen motion polisher.



Fig. 8-38 Photograph of the Momentum,[10] an orbital platen motion polisher. Courtesy of SpeedFam-IPEC, Inc.

defect or particle on the chuck (or the back of the wafer) will cause a thin spot in the film being polished. In addition, the possibility of wafer breakage is

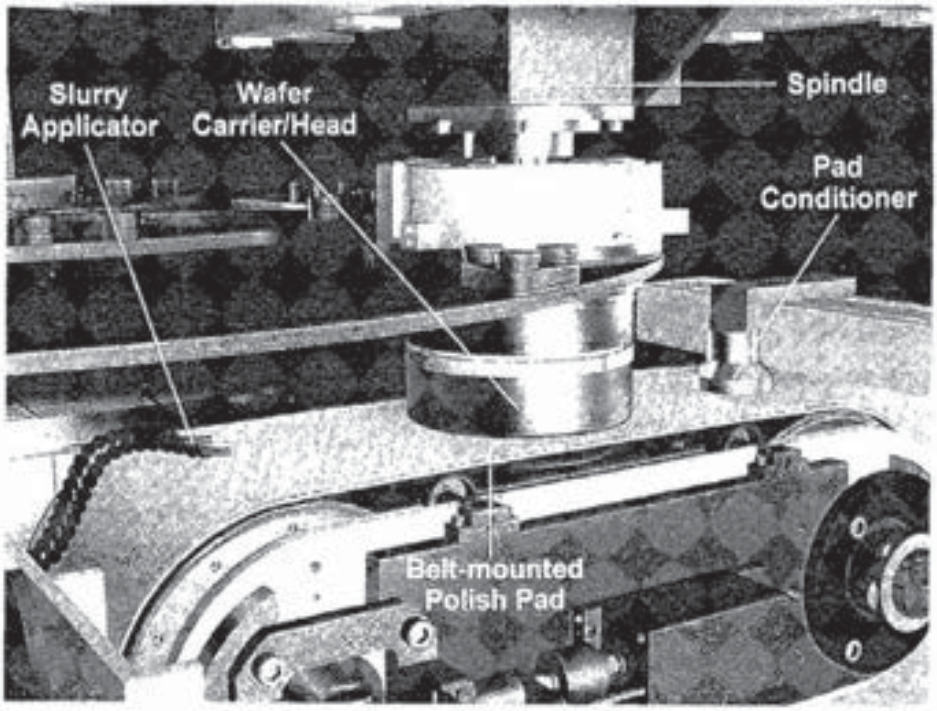


Fig. 8-39 Schematic drawing of the Teres® linear CMP tool. Courtesy Lam Research, Inc.

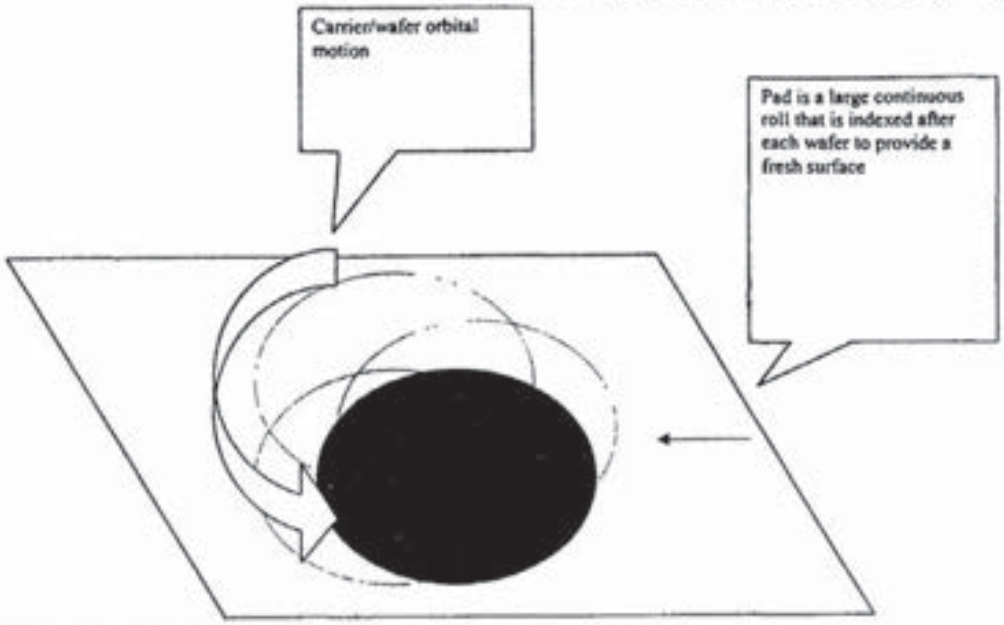


Fig. 8-40 Schematic drawing of orbital carrier-motion with indexing pad.

increased. Early backing films were merely thin foam layers. They did little to lessen the polish-performance-dependency on the variations of wafer-thickness or flatness. In order to more fully compensate for wafer-to-wafer variations in flatness and film-thickness, much development work has been done to improve the wafer-carrier design.

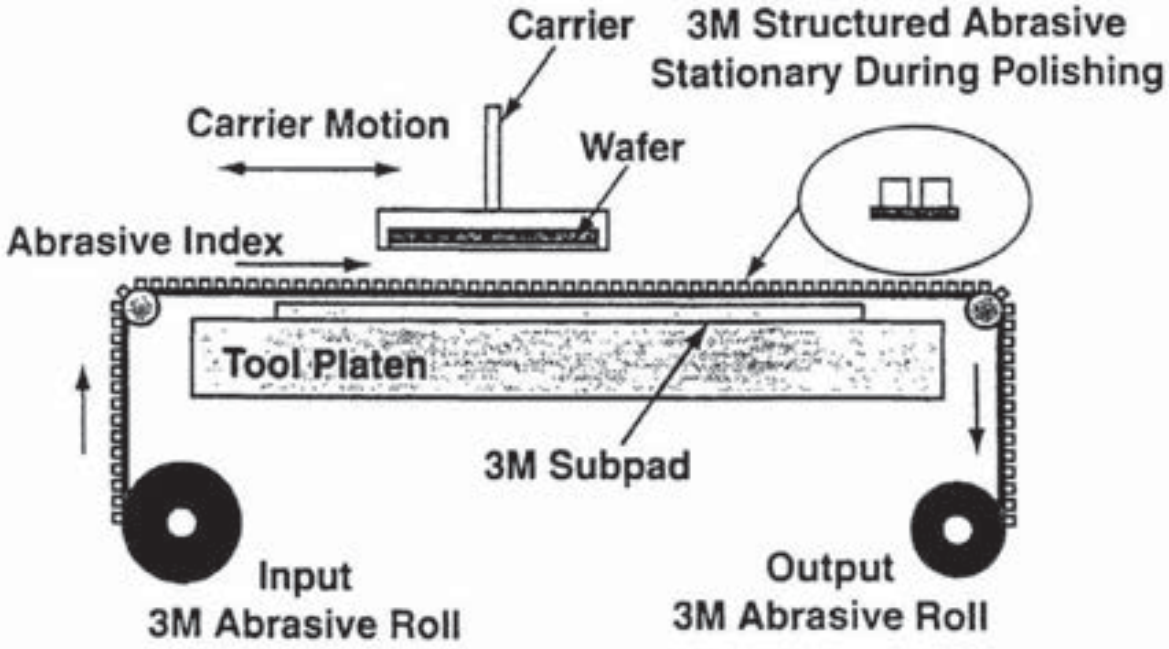


Fig. 8-41 Fixed-table with roll-to-roll pad. Carrier rotates and also executes linear motion.

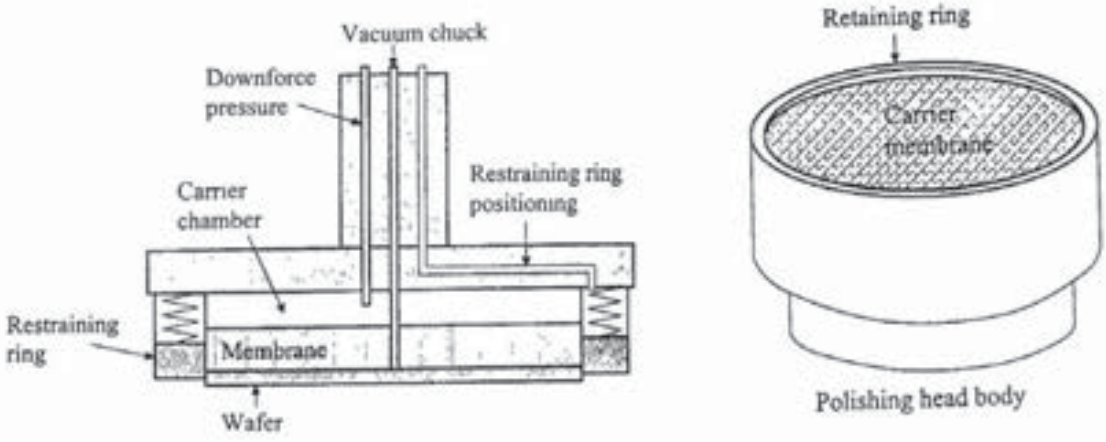


Fig. 8-42 Schematic drawing of the components of a CMP wafer-carrier.

Two key design elements in the improved wafer-carriers are:

1. The retaining-ring;
2. The wafer-leveling mechanism.

The primary function of the *retaining-ring* is to prevent the wafer from slipping out of the pocket when high down-force and high-speeds are used. Wafer retention can be maintained even with an average down-force of about 2-psi if the retaining-ring extends past the mid-thickness of the wafer. To keep the retaining-ring surface from being polished, some wafer-carriers may keep the retaining-ring about 100-microns off the pad surface. Wafer-carriers with such rings are termed *non-contact retaining-ring carriers* (NRRC). In NRRCs there is no pressure applied to the pad by this ring. The other type of retaining-ring

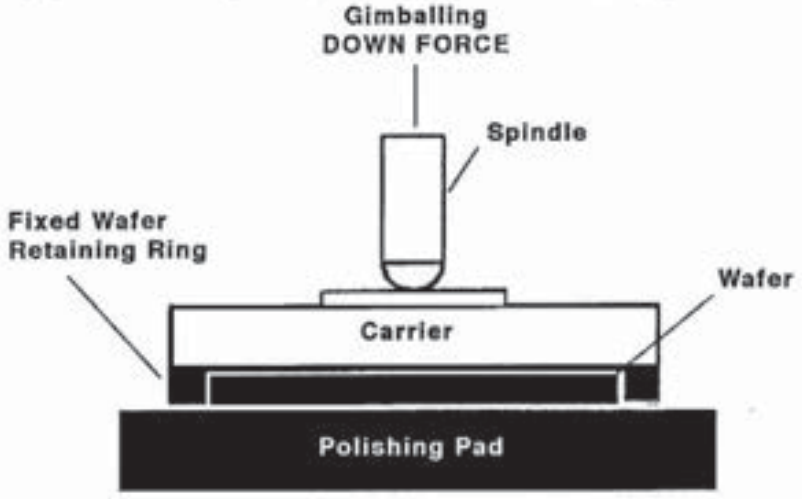


Fig. 8-43 Carrier being subjected to gimballing down force.

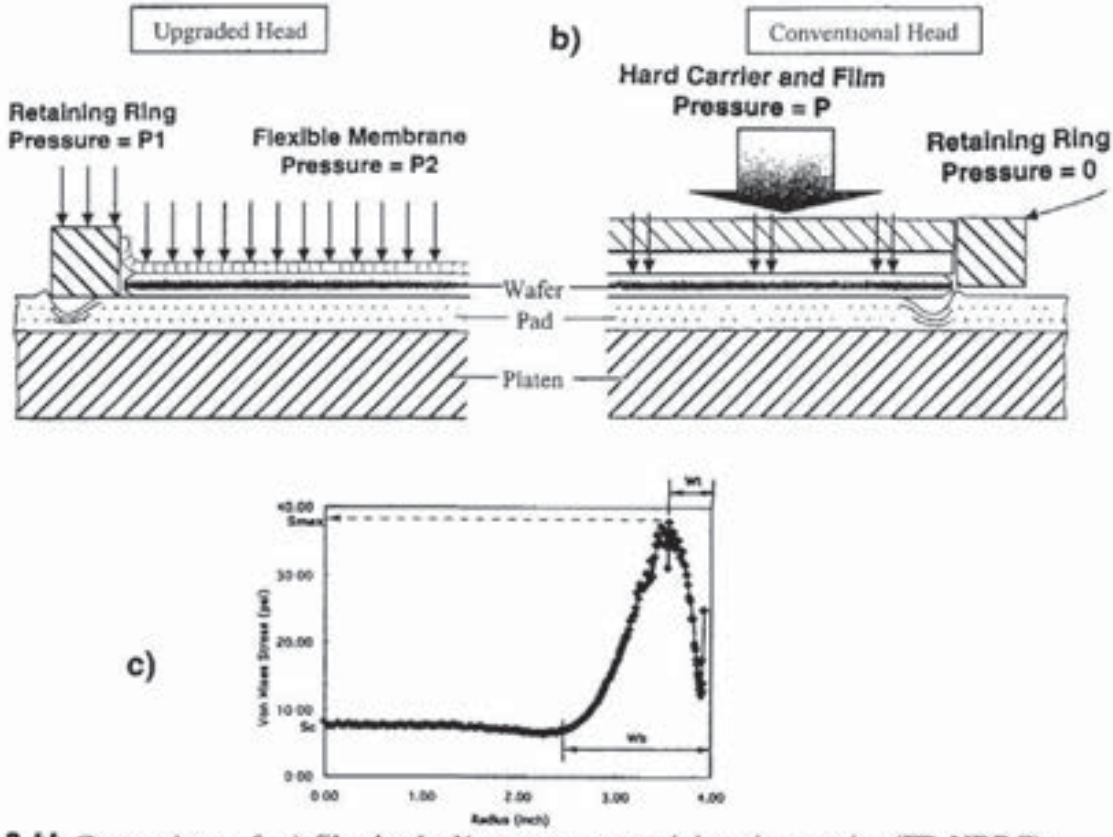


Fig. 8-44 Comparison of: a) film-backed/non-contact retaining-ring carrier (FB-NRRC); and b) bladder-backed/contact retaining-ring carrier (BB-CRRC); c) distribution of pressure profile at wafer edge for FB-NRRC.

design causes the retaining-ring to contact the pad. Hence, carriers with these rings are called *contact retaining-ring carriers* (CRRC). The CRRC also has the capability of varying the down-force on the wide (typically 10-25 mm-wide) bearing surface of the ring. Wafer-carriers used in some polishing tools (e.g., Applied Materials' Mirra® and Strasbaugh's Symphony®) are designed with a retaining-ring that can apply variable pressure to the pad.[28] This allows the pad profile at the wafers edge to be controlled. A smaller edge-exclusion zone on the wafer can therefore be used (2-3 mm).

Wafer leveling by the wafer-carrier is accomplished by backing the wafer with: **1**) a compressible film (i.e., *film back carrier,* FBC); **2**) a pneumatic bladder (*bladder-backed carrier*, BBC); or **3**) a lip-sealed air bearing *(air-backed carrier*, ABC).

The earliest wafer carriers were FBC-NRRC (Fig. 8-44a). Perforations in the backing film and the flat surface plate are used to vacuum-attach the wafer to the carrier during transport. The backing film acts as a cushion to distribute the down

force evenly to the wafer. Early backing films were chemically-impregnated felt materials that had to be applied by hand, but they were only useful for about 25 wafers. Newer backing films (such as the DF200, the WB, and the R200 films) last about 200-800 wafers. (The reason backing films wear out is that they are compressed and de-compressed each time they are used. At some point, the film will lose its ability to recover and takes a "set.") In FBC-NRRCs the pad pressure at the edge of the wafer is not well controlled, and the polishing rate of the film is decreased at the wafer's edge. Hence, a wider edge-exclusion ring of 5-7-mm must be enforced for wafers with such carriers.

The second type of wafer-carrier is the BBC-design (Fig. 8-44b). In the BBC a rubber inner-tube-style bladder is substituted for the backing film. The chamber created by the bladder can be alternately vacuumed or pressurized. When the carrier brings the wafer into contact with the platen, first pressure is applied to the retaining-ring bladder (with a sufficient down force to clamp the wafer within the retaining-ring). Next, the wafer-bladder may be vented or pressurized. Usually the platen is stationary when the wafer-carrier retaining-ring makes contact and the wafer-bladder is vented until the wafer-carrier comes to speed. Then the wafer-bladder is pressurized. Since the wafer "floats" on the bladder bubble, global and local planarity of the wafer surface is accomplished without gimbaling the wafer-carrier to the head of the tool.

The air-backed wafer-carrier (ABC) functions much like the BBC. However, instead of a bladder to contain the air pressure (or vacuum) behind the wafer, it has a compliant lip-seal around the backside edge of the wafer. However, the seal is not easy to maintain, and slurry particles can enter and contaminate the backside of the wafer. As a result, ABC types have not yet been commercialized.

8.6.2.1 CMP Edge-Effect

The removal-rate of material by CMP should ideally be uniform across the wafer (i.e., the WIWNU should = 0). However, for various reasons, non-uniformity of material removal within a wafer is generally observed. One of the most common forms of WIWNU is the *CMP edge-effect*, in which the removal-rate is much smaller at the edge of the wafer, leaving an edge-bead of material at the conclusion of the CMP process. Figure 8-45 shows two typical thickness profiles of wafers polished by two different polishers, along a wafer-diameter. In both cases, an edge-bead, 100-200-nm high, exists after the polishing step. This edge-bead will generally render devices fabricated in that area inoperable. As a result, devices fabricated within a region near the edge of the wafer are considered to be excluded as candidates for working die, and this ring near the edge of the wafer

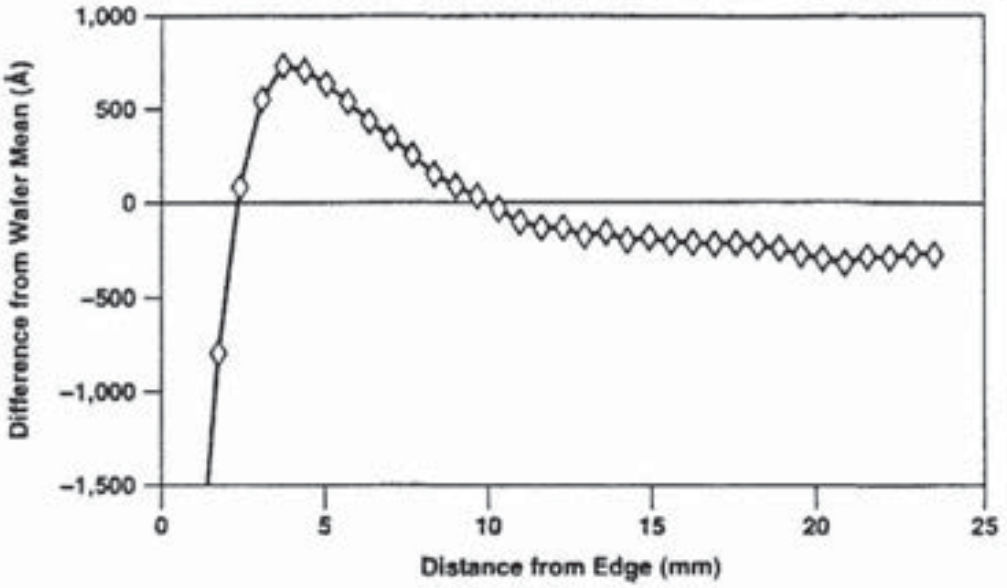


Fig. 8-45 Edge effect observed in CMP.

is termed the *exclusion ring* (Fig. 8-46). In current polishers, the exclusion ring is 3-mm wide. However, as wafer sizes are increased, the number of die at the wafer's edges keeps increasing (Table 8-2).[29] In order to obtain more yielding die per wafer, methods to further reduce the edge exclusion ring are being pursued.

The edge-bead-effect is attributed to the pad-deformation-profile and the pressure-distribution near the edge of the wafer. As a result, this effect can be significantly reduced by optimizing the pressure distribution at the edge of the wafer, primarily through the use of contact-retaining-ring-carriers (CRRC). The bulk physical properties of the polishing pad (such as its hardness, flexibility, compressibility, energy-absorption, and rebound) also all influence the width of

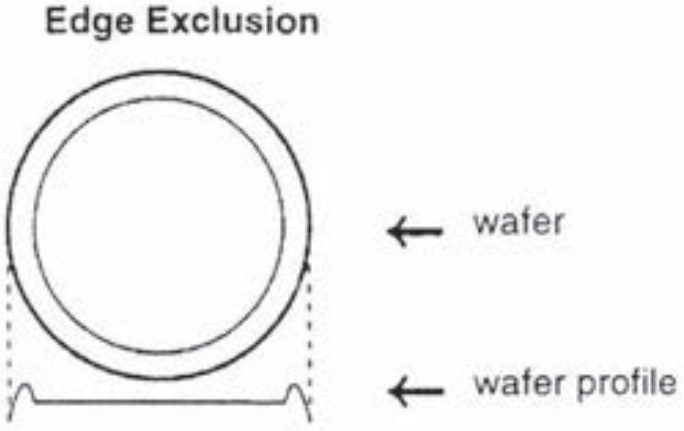


Fig. 8-46 Exclusion ring.

TABLE 8-2 Edge Die Count versus Wafer Diameter

Wafer diameter	Total wafer die count	Edge die count	Edge die percentage	Max edge revenue ($B)
300 mm	515	107	21%	2.76
200 mm	217	69	32%	1.78
150 mm	115	48	42%	1.24
127 mm	71	38	53%	1.00
100 mm	43	26	60%	0.68
76 mm	21	16	76%	0.43

the exclusion ring. Pad materials that provide the hardness necessary for planarization (but remain compliant enough to transfer dynamic contact-pressure more evenly), will deliver improved edge-exclusion. New materials for polishing pads and sub-pads have been designed to improve the uniformity of contact.[30]

8.6.3 CMP Consumables (Slurries)

Slurries consist of small, abrasive particles (generally inorganic oxides) of specific shape and size (typically 10–100-nm) suspended in an aqueous solution. The abrasive particles have roughly the same hardness as the film being polished. Reagent chemicals, such as acids, bases, surfactants, buffers, corrosion inhibitors, and complexing agents, are also added to the solution, depending upon the material to be polished. The slurry parameters that impact the polishing-rate include the chemical composition of the slurry solution (and its pH) and the concentration (in wt %) of the solid-particles in the slurry. Also important are the shape, size, and size distribution of the slurry-particles. An automatic slurry feeding system is used to ensure uniform wetting of the polishing pad and proper delivery of the slurry. Slurries are marketed by a number of suppliers, the largest of which is Cabot Corp. More slurry vendors are entering the market, including Rodel Inc., Fujimi Inc., Olin Microelectronic Materials, Praxair Surface Technologies, and EKC Technology, Inc.

The goal for each CMP process is to find a slurry that produces high removal-rates, good-planarity, local-film-uniformity, and high-selectivity. Also important are the ease with which the slurry can be cleaned from the wafer, the aggressiveness with which it attacks the polishing equipment, its storage stability, and the cost of disposing of the slurry waste. From a manufacturability perspective, consistency is a critical issue. That is, slurries have tended to exhibit quite a lot of variability from batch-to-batch. In the future, suppliers will be

expected to deliver slurry that has the same performance consistently. If this can be done, the need for monitor wafers will be significantly reduced, allowing the cost of the CMP process to be decreased.

The shelf-life of slurry is also an important characteristic. This determines how long a slurry can be stored before use after manufacture. Some slurries need to be kept within a specific temperature-range during shipping and storage to maintain their performance. The shorter the shelf-life of the slurry, the narrower the window in which it can be used, and inventory-control becomes important. An ideal shelf life for slurries is at least one year.

As CMP technology has evolved, slurry technology has therefore also had to progress to keep pace with the demanding requirements of the industry. Today's CMP slurries are increasingly tailored for specific process applications (like STI or Cu interconnects) as well as for specific polishing tools or other consumables (i.e., pads). Some of these issues have been discussed in Sect. 8.5.2, and others will be addressed in Sect. 8.11. The remaining slurry topics are considered here.

One other aspect of slurry management should also be pointed out. Slurries tend to dry rapidly on a wafer after leaving the wet environment of the polishing table. Once a slurry dries, it is very difficult (if not impossible), to remove all the slurry particles from a wafer surface. To prevent this from occurring, wafers are

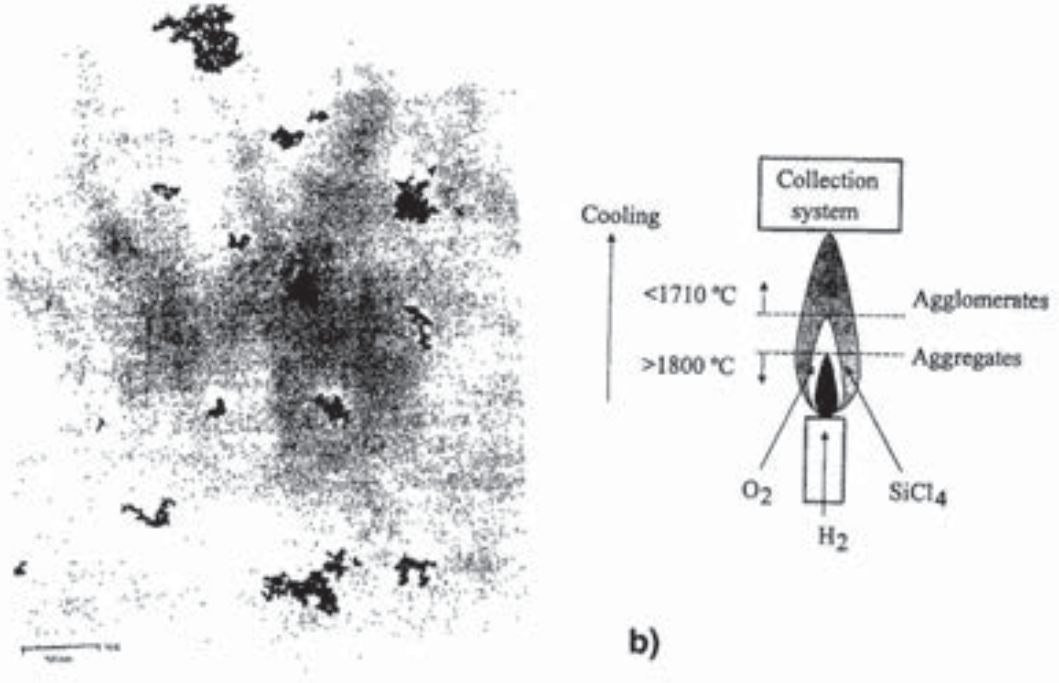


Fig. 8-47 a) SEM of fumed-silica slurry-particles. Courtesy Rodel; b) Fumed-silica formation.

rinsed while still on the pad, and then quickly transported into wet-cleaning tanks to remove the slurry. In fabs that do not use polishing tools with integrated cleaning systems, logistical difficulties may delay the transport of wafers from the polisher to the cleaning system, inviting the above problem.

8.6.3.1 Slurries for Oxide-CMP: Slurries for oxide-CMP traditionally consisted of colloidal-silica or fumed-silica suspended in an alkaline solution (typically potassium-hydroxide [KOH] or ammonium-hydroxide [NH_4OH], diluted in DI-water). The first generation of slurries for oxide-CMP contained fumed-silica abrasives, and they still retain a large fraction of the market share. Fumed-silica is manufactured as a dry powder by the oxidation of silicon tetrachloride ($SiCl_4$) at high temperature in a flame (see Fig. 8-47). The powder must be dispersed and blended into the final slurry. The fact that the fumed-silica must be mixed is inconvenient, and adds cost. The average selling price of fumed-silica slurries in 1998 was \$10-\$15 per gallon. Figure 8-47 also shows a SEM of fumed-silica slurry particles.[31]

Colloidal-silica slurries have entered the market as an alternative to fumed-silica. Colloidal-silica is manufactured by the hydrolysis of sodium-silicate (or other silicates). This is followed by an ion-exchange to reduce the sodium levels for IC applications. (Early on, colloidal-silica was avoided due to the possibility of sodium contamination, but improvements in this aspect of colloidal-silica processing have opened the door for its adoption.) Colloidal-silica is prepared directly in solution, thus avoiding most of the handling problems of fumed-silica. Its microstructure is a distribution of spherical particles (see Fig. 8-48).[32] Colloidal-silica slurries also tend to be more stable and less prone to agglomeration than fumed-silica slurries. In addition, their more uniform size

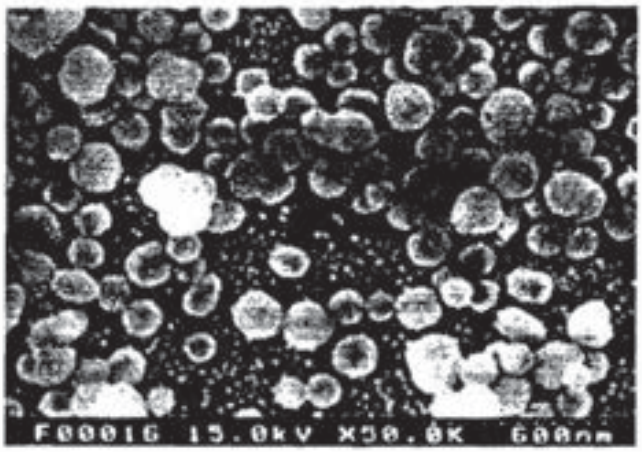


Fig. 8-48 SEM of colloidal-silica slurry particles. Courtesy of Fujimi Corporation

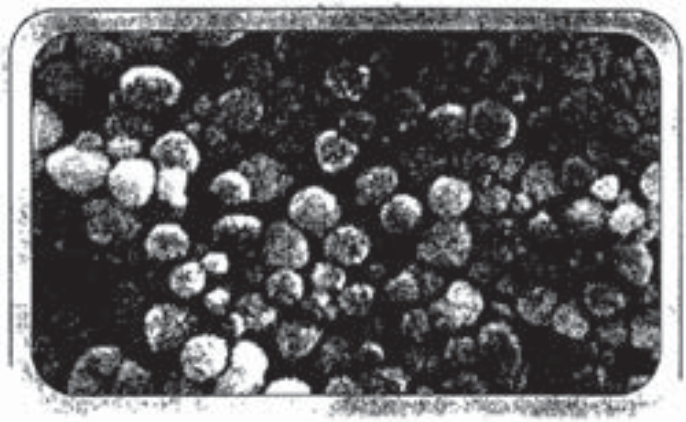


Fig. 8-49 SEM of cerium oxide slurry particles. Courtesy of Fujimi Corporation.

distribution makes it less likely that they will cause microscratches. The cost of colloidal-silica slurries is about 20% higher than that of fumed-silica slurries ($18 versus $15 per gallon). The use of colloidal-silica slurries has rapidly increased (replacing fumed-silica slurries), and they find use in STI, ILD, and polysilicon CMP processes.

The solids content of oxide-CMP slurries is 5-35-wt%, and the pH is between 9.0–11.0. By controlling pH within this range, agglomeration is minimized. CMP slurries generally contain few chemical additives. As mentioned above, these few typically consist of: **1**) an alkaline chemical to provide colloidal stability (i.e., to prevent slurry-particle-agglomeration) and to increase oxide removal rates; and **2**) a buffering agent (used to keep the pH constant throughout the slurry volume and over time). The polishing rate of oxide increases with: **1**) increasing pH,; **2**) increasing slurry particle concentration; and **3**) increasing particle size.

Slurries for oxide-CMP based on other abrasives than silica have also been introduced in recent years. Ceria-slurries (CeO_2-based) have been formulated for STI and ILD applications, although ceria is difficult to manufacture at small particle sizes. A self-stopping, dishing-free process using a ceria-slurry was reported by Nojo.[33] Ceria-slurries exhibit a higher removal rate at a much lower pH than silica-slurries (e.g., at a pH of 6 to 7). Ceria-slurries have been adopted for STI-CMP because of their high-selectivity to oxide over nitride (as high as 250:1). Ceria particles can also be blended with silica particles in a slurry. The addition of silica helps keep ceria particles in suspension. In 1998, the average selling price of ceria-slurries was $18 to $20 per gallon. Figure 8-49 shows cerium oxide (CeO_2) particles.[34]

Zirconia-based slurries (ZrO_2) have been investigated for use in low-*k*-dielectric-CMP, especially the CMP of non-fluorinated polyarylene ethers (e.g.,

FLARE™, from Honeywell, see Chap. 14). Zirconia is harder than silica but softer than alumina. Manganese-oxide slurries (MnO_2) have also been studied for the direct polishing of low-*k* dielectrics (and for the CMP of W). It was reported that the deep scratches formed with CMP using a conventional silica-based slurry were reduced from a 20-nm depth to a 2-nm depth by the use of a MnO_2-based slurry. Residual MnO_2 slurry particles can also be removed using H_2O_2 with a pH of 6, which does not attack the fluoro-carbon low-*k* material being polished (e.g., a poly-AE layer, FLARE®).[35]

8.6.3.2 Slurries for Metal-CMP: Slurries for metal-CMP applications typically have more constituents than do oxide slurries. Besides the *oxidizer* (e.g., ferric-nitrate and hydrogen-peroxide), and a buffer agent, they may contain: **1**) a *surfactant* (to assist in keeping the slurry particles dispersed, see Sect. 8.7.1.4); **2**) a *complexing agent* (used to facilitate the removal of metal from surfaces, especially in the case of copper); **3**) *corrosion inhibitors* (e.g., benzotriazole [BTA] in Cu slurries to permit the use of acidic slurries); and **4**) *additives to change the selectivity to other films*. The oxidizer is often separated from the abrasive for reasons of colloidal stability or shelf life. In the year 2000 time-frame, choosing and balancing all of these constituents still makes the formulation of metal-CMP slurries as much an art as a science.

Metal-CMP slurries typically use alumina particles suspended in various liquids (depending on the specific metal being polished). Alumina slurries are preferred because they exhibit higher metal-removal-rates and higher selectivity to the oxide ILD-layers (> 20:1, metal-to-oxide). The alumina particles may be fabricated either by a fumed process similar to silica, or by precipitation from aluminum hydroxide.

The pH range of metal-CMP slurries is low (e.g., pH = 1.0-5.0 for W-CMP alumina slurries, and pH = 3.0- 9.0 for Cu-CMP alumina slurries). As noted in the previous section, silica particles in oxide-slurries can be kept well dispersed for relatively long periods of time by controlling the pH of the slurry (giving them a long shelf-life). However, alumina-based slurries have traditionally exhibited the drawback of poor colloidal stability. This causes slurry particle agglomerates to form. Such clumps produce unacceptable microscratches in the films being polished. Recently, improvements have been made in the colloidal stability of alumina based slurries.[36] Nevertheless, to overcome this problem from another direction development of slurries based on other particle types (e.g., cerium oxide and zirconium oxide particles) for metal-CMP has been pursued. In 1998, the cost of alumina slurries was between $25 and $30 per gallon.

8.6.3.3 Multi-Step Slurries: When carrying out a metal-CMP process (e.g., W or Cu) two materials must be polished (i.e., the main metal, and the underlying barrier [or glue] layer). This CMP-step can be carried out with a process that uses one slurry, or two slurries. While the one-slurry process is simpler, it has not been possible to restrict the problems of dishing and erosion to the necessary degree with such a process. Hence, the two-slurry approach has been adopted. That is, for Cu-CMP, the first slurry removes the bulk of the Cu in the field areas, while the second removes the barrier (e.g., Ta or TaN) from the field areas. These two-step CMP processes significantly reduce the copper dishing and erosion effects. (Generally, the formulations of these slurries still remains proprietary). Use of a CMP-tool with multiple-platens also facilitates the manufacturability of such two-step CMP processes.

8.6.4 Slurry Distribution Systems

Slurry management (i.e., mixing, storage, distribution, and dispensing) is another key issue of a CMP process. Reliable *slurry distribution systems* (*SDS*) are an increasingly important component of the overall CMP-equipment-set. It is critical that the user of a CMP system view the SDS as an integral part of the CMP manufacturing process, and treat it with as much care as a polisher is given. This will greatly minimize process variations, ensure smooth functioning of manufacturing, and reduce unscheduled downtime.

Slurry management is challenging for a number of reasons.[37] Precise mixing and consistent batch blends are necessary to achieve the repeatability demanded by process engineers. The slurry consists of particles in a liquid, and they must be kept suspended in it in a homogeneous state. However, these particles will agglomerate into clumps unless steps are taken to suppress this tendency. (The main source of surface deformations in CMP is generally believed to be the presence of undetected coarser-size particles in the CMP slurry.)[38] The slurry-pH plays a role in the dispersion of silica particles. As is described in more detail in the section *CMP Cleaning*, at a pH above 7.5 the silica particles acquire sufficient surface charge to generate electrostatic repulsion, and this effectively keeps the slurry well dispersed.[39] (Above a pH of 10.7, however, the particles dissolve and form a silicate.) In addition, from the time the raw slurry is introduced into the SDS it is also critical that a blanket of humidified air be provided over the slurry to prevent agglomeration in the dispense system. Finally, slurries for Cu- and W-CMP require constant re-circulation to maintain particle suspension. (It is typically recommended that a minimum flow-rate of 1-m/sec be used for re-circ-

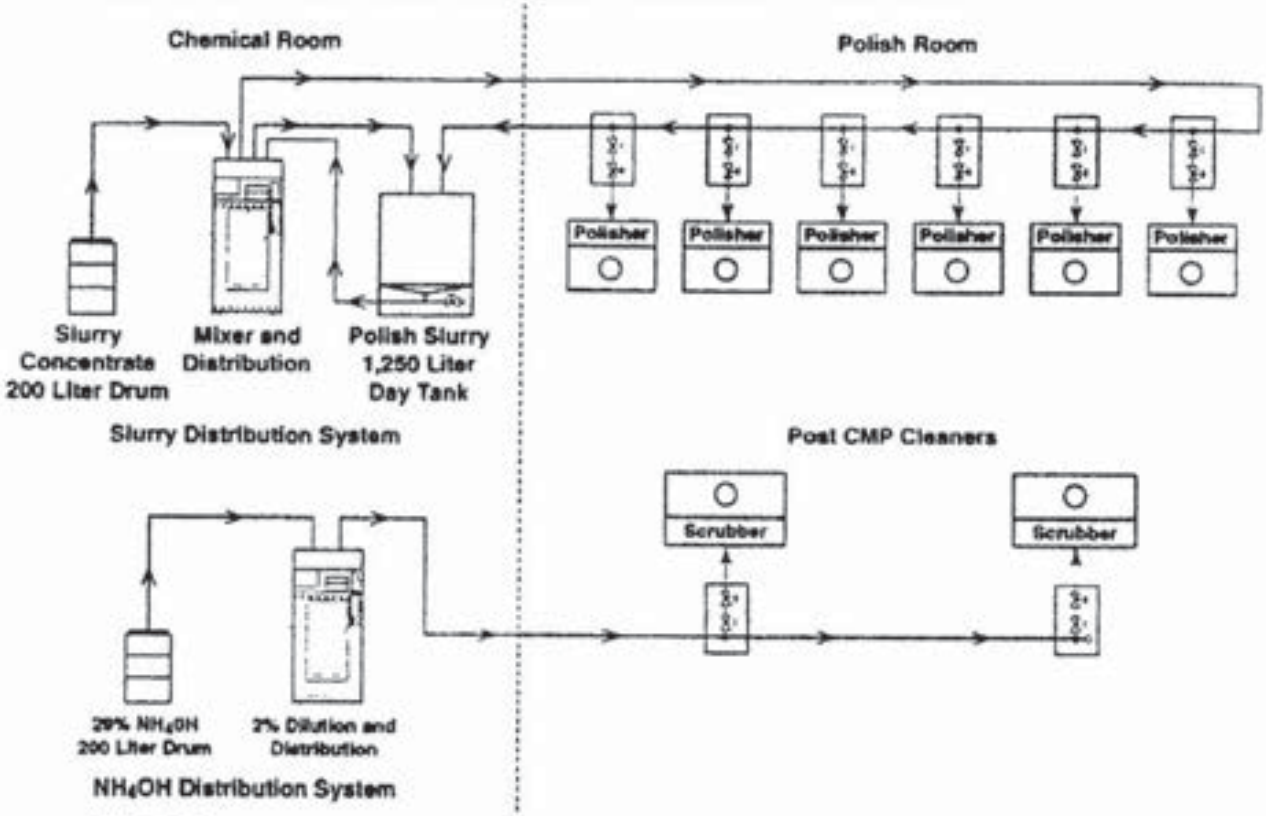


Fig. 8-50 Slurry distribution system.

ulation). Slurries have a finite shelf-life (after which the agglomerate-level becomes excessive). This shelf-life ranges from less than one week to about one month. Filtering is therefore necessary to prevent any agglomerates (which might have formed during that time), from reaching the wafer being polished.

Slurry handling in an SDS is a delicate task and care must be exercised to prevent "damaging" the slurry. That is, the slurry must not be subjected to excessive shear (which could cause the slurry to agglomerate), or to insufficient shear (which would permit slurry settling). The SDS must also be regularly maintained. Such maintenance involves periodic cleaning out and flushing of the system (as well as the polisher).

The SDS must continuously provide slurry to a number of polishers in a sustained manufacturing production environment (i.e., 24 hours a day, seven days a week, and in high volumes). The pH and solids-content of the slurry must be monitored in-line to ensure consistency. Furthermore, since slurries are abrasive particles contained in chemically aggressive liquids, they present a harsh environment to the components of the distribution system. That is, pumps, piping, valves, and measurement apparatus in contact with slurry, can be stressed to the point of failure in a short period of time. Nevertheless, bulk slurry

distribution systems have been designed to handle these conditions. These sys-systems are sophisticated pieces of equipment, far more advanced than other bulk-delivery systems. Figure 8-50 shows the schematic layout of an SDS, and Fig. 8-51 shows the actual pieces of equipment in an SDS. Such SDS systems are available from vendors including: Mega Systems & Chemicals and FSI International. These vendors supply detailed recommendations regarding design and maintenance of SDS, piping, mixing, and dilution procedures, filtration of the slurry, filter changeouts, and re-circulation of the slurry.

8.6.5 Environment, Safety, and Handling Issues of CMP Slurries

As CMP further moves into large scale IC production, the disposal of spent slurry, CMP process by-products, and metal contaminants (i.e., material removed from the wafer surface) become important issues.

Slurry also always contains aggressive chemicals. In some cases these chemicals are highly toxic. Furthermore, the material removed by a CMP process ends up in the expended slurry and the rinse water. Some IC fabs estimate that CMP processes may come to consume 30-40% of the total water utilized in IC manufacturing. The issue of purification and disposal of this waste water (or

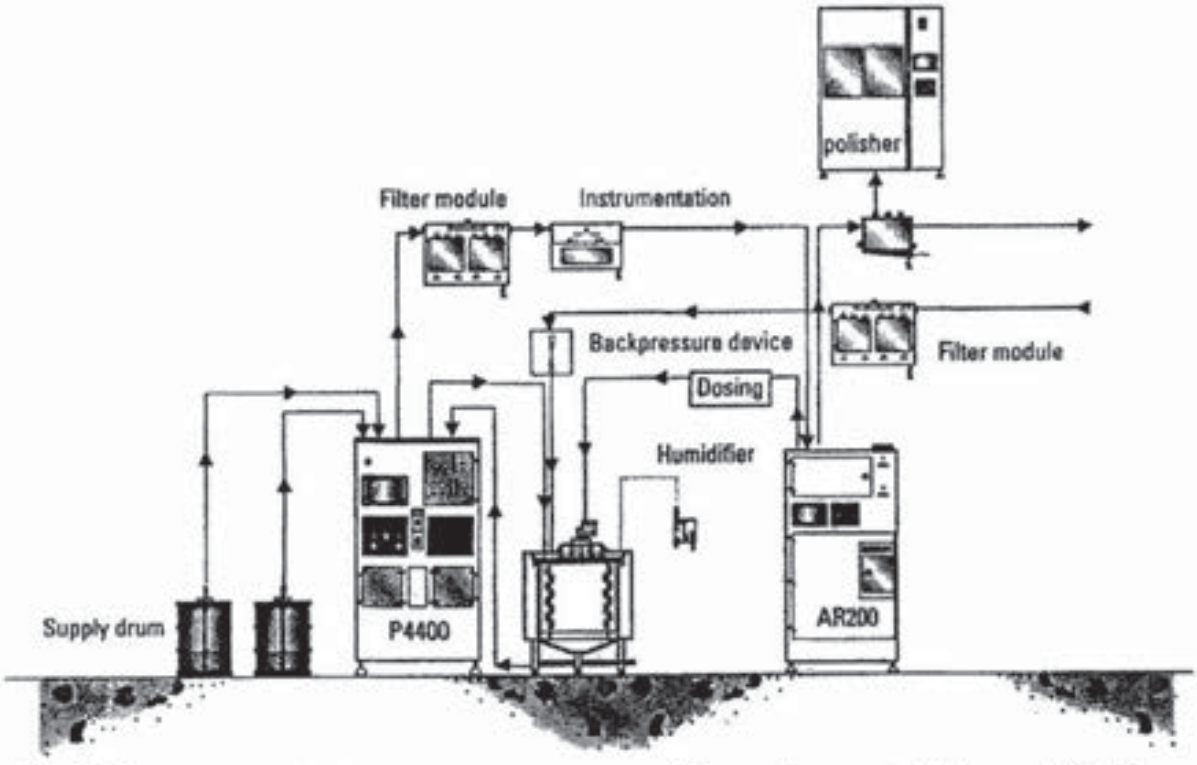


Fig. 8-51 Schematic of a slurry management-system with metrology, auto-dosing and filtration.

recycling), becomes more critical as new materials such as copper, gold, and tantalum become adopted in IC manufacturing. For example, copper is a known toxin in water. Therefore, fabs require waste-water treatment systems to recycle the drain water of CMP processes. The most common limits for waste streams are those for Total Suspended Solids (TSS), heavy metals (like copper, chromium, zinc, and nickel), pH, and total organic content (TOC).[40]

The slurry particles are suspended particles in a solution, and hence they can be removed from the waste water by a process of *flocculation* (i.e., a process by which the dispersed slurry particles are caused to clump together to form larger clusters of matter [i.e., to flock]) followed by filtration. Flocculation involves adding a flocculent and adjusting the pH to a level where the flocculent works best. The slurry particles then agglomerate and settle out of suspension. After the solids are separated, the next step involves ion removal. Either ion-exchange or reverse-osmosis may be used to do this. Note that as slurry volumes go up, they will generate huge amounts of solid waste, estimated at 100,000-400,000 pounds per year, per typical fab, and this solid waste must also be disposed of safely. A decrease in slurry consumption by the CMP-processes will be needed to reduce this waste-disposal burden.

Some companies are also examining the issue of reusing the water from a polisher. That is, while a polisher is in an idle (or pad-wet) mode, there is a continuous stream of DI-water that is used to keep all parts of the polisher wet and reduce the possibility of formation of dry particles. This water generally retains a high purity level and offers the potential of recycle and reuse, thus reducing the waste-treatment-costs for CMP.

One suggested method of reducing slurry consumption is to re-cycle the slurry. In one study of re-using slurry, KOH-based fumed-silica slurry was used to polish a PECVD-TEOS film. The slurry was then captured after being flushed from the CMP tool. By adding new slurry and adjusting the pH to compensate for the decreased solids content and pH, the slurry appeared to be capable of re-use up to six times, while maintaining roughly equal removal rates.[41] In a more comprehensive slurry-recycling process, the slurry-waste would be treated by the following processes: filtration, oxidizer-replenishment, removal of metals, and replenishment of buffers and other necessary additives. The task of slurry re-cycling (and eventual disposal) could be carried out by the slurry manufacturer. On the other hand, it has also been discussed that the slurry could be leased by the end user, and that it would be sent back to the slurry manufacturer after use for regeneration. The concept of slurry re-use is fraught with a number of

difficult challenges, and is today still in its infancy. The issues of co-operation and development between the end-users and suppliers still need to be addressed before this process becomes an accepted procedure.

8.6.6 CMP Consumables (Polishing Pads)

Pads and slurries are the consumables of a CMP process. Historically, the polishing tools were considered to be the technology's major contributors. But consumables are now being recognized as being equally important. With a new group of consumables being made available, some customers select the slurries first, then review their compatibility with existing tools, and based on the results, select a polishing tool. The importance of CMP consumables is expected to increase with the advent of Cu-CMP and low-*k*-CMP. The general characteristics of polishing pads are described here.

Polishing pads are made of either: **1**) polyurethane-impregnated felts; or **2**) closed-cell cast and sliced polyurethane (with filler material to control hardness). Urethane is utilized because urethane chemistry allows the pad characteristics to be tailored to meet specific mechanical property needs. Table 8-3 lists the important properties of five polishing pads (all manufactured by Rodel). A thorough discussion of the mechanical properties and the surface roughness of CMP pads is found in pp. 68-83 of Ref. 42.

The pads made from impregnated felts are more porous and can carry a larger volume of water or slurry, due to the basic fiber structure of felt. The Rodel Suba

TABLE 8-3 Rodel Polyurethane Pad Properties

Pad	Specific Gravity	Compressibility	Hardness	Type
Politex Supreme	*Not available*	High	Soft	Micro porous felt
Suba IV	0.3	15%	61 (Shore A)	Urethane impregnated polyester felt
Suba 500	0.34	7%	72 (Shore A)	Urethane impregnated polyester felt
IC-1000	0.63–0.85	2.25%	57 (Shore D)	Rigid micro porous cast polyurethane
IC-1400	Top layer IC-1000, bottom layer foam	0.7%–7%	Flexible	Equal to stacked IC1000/Suba IV

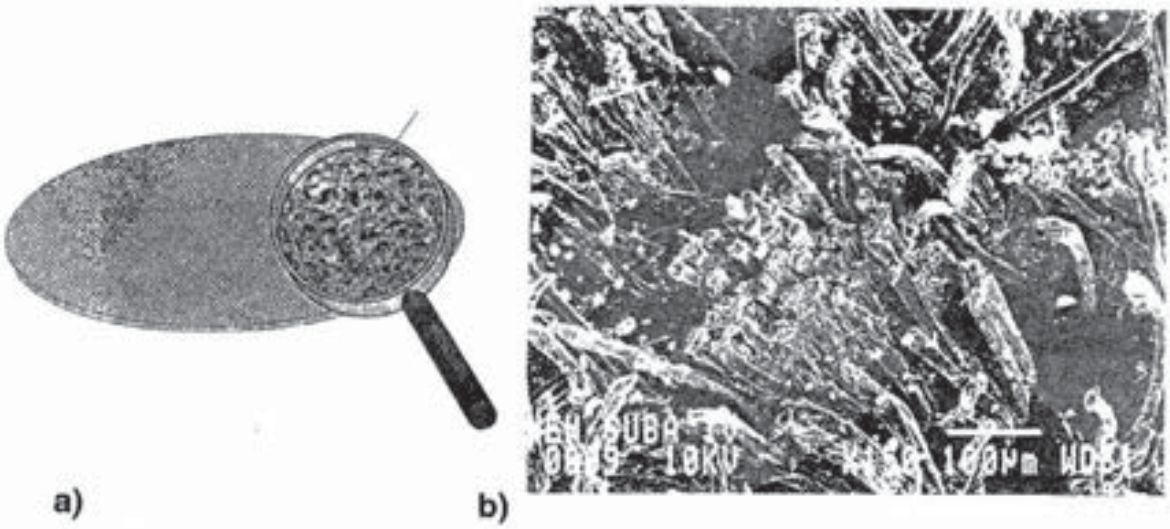


Fig. 8-52 a) Schematic of polishing pad; b) SEM of a new Rodel Suba IV® Pad. Courtesy of Rodel Inc.

IV pads are made of polyurethane impregnated polyethylene felt (Fig. 8-52), which shows a SEM of the surface of a new Suba IV pad.[43] They are classified as *soft-polish pads*, and exhibit greater compressibility than the closed-cell polyurethane types. The Rodel Politex pad is also made from a flexible felt-like material, and it is classified as a *finishing pad*, used to buff wafers with water and other chemicals to remove particles and reduce scratch defects.

The Rodel IC-1000 pads are micro-porous, filled, cast pads that have low compressibility (Fig. 8-53).[44] Pores make up about 30% of the volume of the pad, and the average diameter of a pore is about 30-nm. Such micro-porous, cast pads are classified as *hard-polish pads*. The IC-1000 pad has been the most widely used pad for oxide-CMP processes. The Rodel IC-1400 is a hybrid pad. While there is no consensus, it is generally agreed that the hardness and porosity

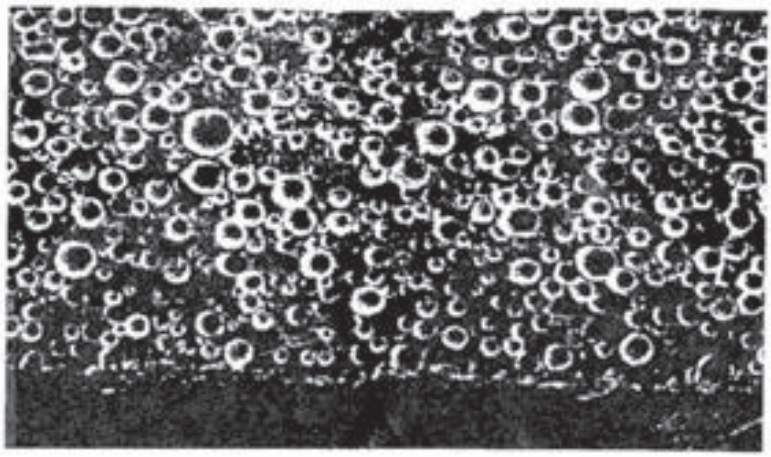


Fig. 8-52 SEM of a new Rodel 1000 IC Pad. Courtesy of Rodel Inc.

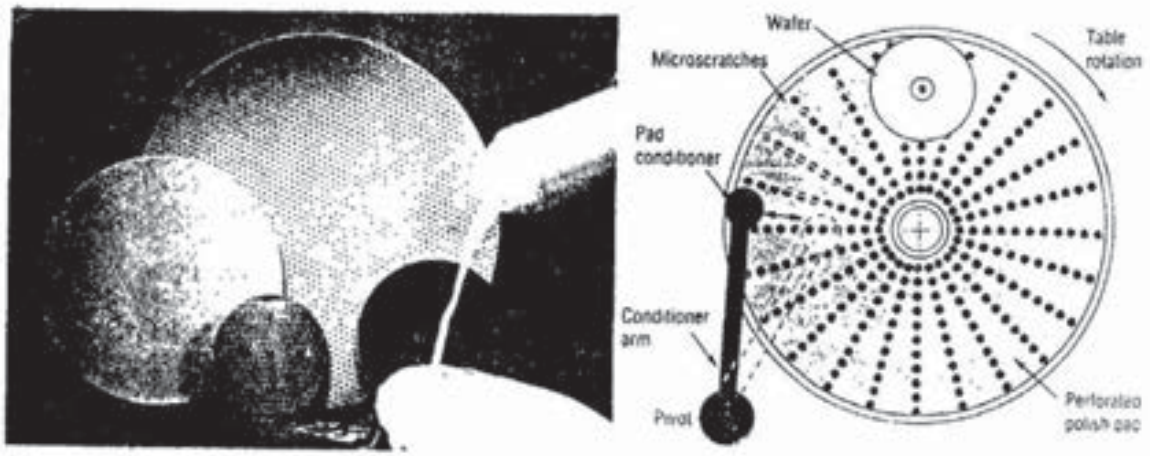


Fig. 8-54 Photograph of polishing pads, showing pad perforations. Courtesy of Rodel Inc.

of the pad are important parameters and that the pad plays a crucial (but still poorly understood) role in determining the final polish results. However, there is no single physical property of a polish pad that is known to lead to good quality for all CMP tools and consumable sets. Polishing pads are supplied by a number of vendors, with Rodel Corp. having the largest market share. Other suppliers of pads include Thomas West, Freudenberg, Fujimi, and Teijin.

In many CMP applications, two pads are simultaneously used, because a hard pad gives better local (within die) planarity, but a softer pad gives better uniformity of material removal across the entire wafer. Using a combination of two pads provides a compromise between these extremes. For example a hard IC-1000 pad is mounted onto a softer Suba-IV pad to form a stacked pad. On the other hand, the IC-1400 is a single molded pad. That is, it is an IC-1000 pad molded to a more porous backing pad that simulates the compressibility of the Suba-IV pad.

As shown in Figures 8-54 and 8-55, the surface of the pad is either cut with concentric-grooves or contains perforations 1-mm-wide and 250-μm-deep punched into the pad. These grooves (or perforations) form channels to help transport the slurry between the pad and wafer. Pores present on the pad surface also aid in slurry transport.

The desired goal of cost-of-ownership for CMP pads has been identified as less than \$0.05 per wafer. The cost of closed-cell polyurethane pads is high (\$0.06-\$0.10/cm^2). Thus, for a 200-mm CMP rotary tool (which requires a 300-mm diameter pad) the initial cost for the pad is about \$50. For a 300-mm CMP tool, the pad must have a diameter of 510-550-mm, and the cost of the pad is about \$125. To achieve a CoO for the pad of less than \$0.05 per wafer, the

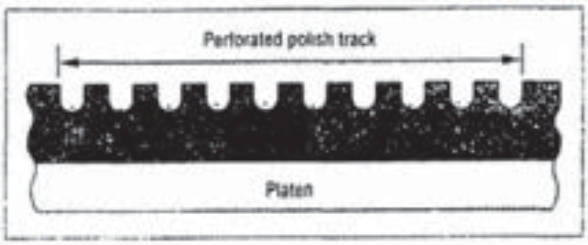


Fig. 8-55 Polishing pad cross-section, showing grooves in pad surface.

smaller pad (i.e., for 200-mm wafers) must polish 800 wafers, and the pad for the 300-mm CMP tool must polish 2500 wafers before being discarded.[45]

8.6.7 Pad Conditioners

One crucial aspect of CMP is the state of the surface of the pad. That is, surface roughness and porosity determine: **1**) slurry transport to the wafer surface; **2**) material transport away from the wafer surface; and **3**) the contact area of the pad to the wafer surface. As a result, unless the surface of a pad is continuously maintained in the optimal state of roughness and porosity, removal-rates can vary with time. In fact, a decrease in removal-rate has been most commonly observed (approximately logarithmic with time).[46] For example, starting with a new pad, the removal-rate can drop from 210-nm/min to 75-nm/min after only 50 wafers have been polished. This effect occurs because during the polishing process the surface of the pad rapidly undergoes plastic deformation so that the surface becomes smoother and the pores become filled with pad material (Fig. 8-56). This phenomenon has been termed *glazing*. Such glazing is undesirable, since as the pores of the pad become closed, slurry delivery is decreased to the wafer surface, and reduced polishing rates result. In addition, large waste-particles can become embedded in the pad and cause scratches.



Fig. 8-56 Surface of a glazed pad (Rodel's Suba IV® pad) after polishing.

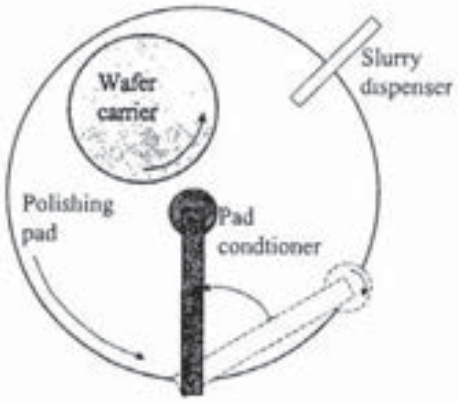


Fig. 8-57 Schematic drawing of pad conditioning. Reprinted with permission of Solid State Technology. Published by PennWell.

A consistent (and rapid) removal rate is necessary for production-worthy CMP processes. This can be achieved by a process that refreshes the surface of the pad, and this technique is called *pad conditioning*. Figures 8-57 and 8-58 show a typical conditioning process in which the pad-conditioner (a disk whose surface is embedded with a diamond grit) sweeps back-and-forth across the rotating pad to restore the pad to its required surface planarity and surface roughness. The pad-conditioning process also opens the pores of the pad by forming micro-scratches on the pad surface. An example of what the pad-conditioning process accomplishes is shown in Fig. 8-59, which contains SEM micrographs of the

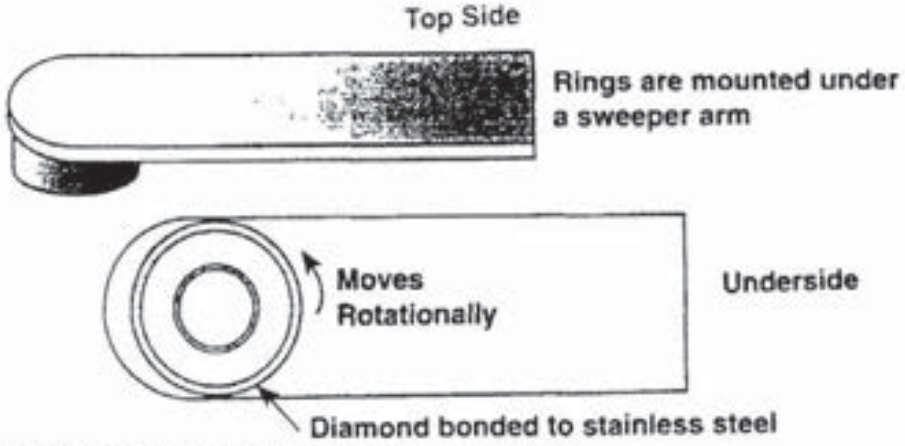


Fig. 8-58 Diamond embedded ring pad conditioner.

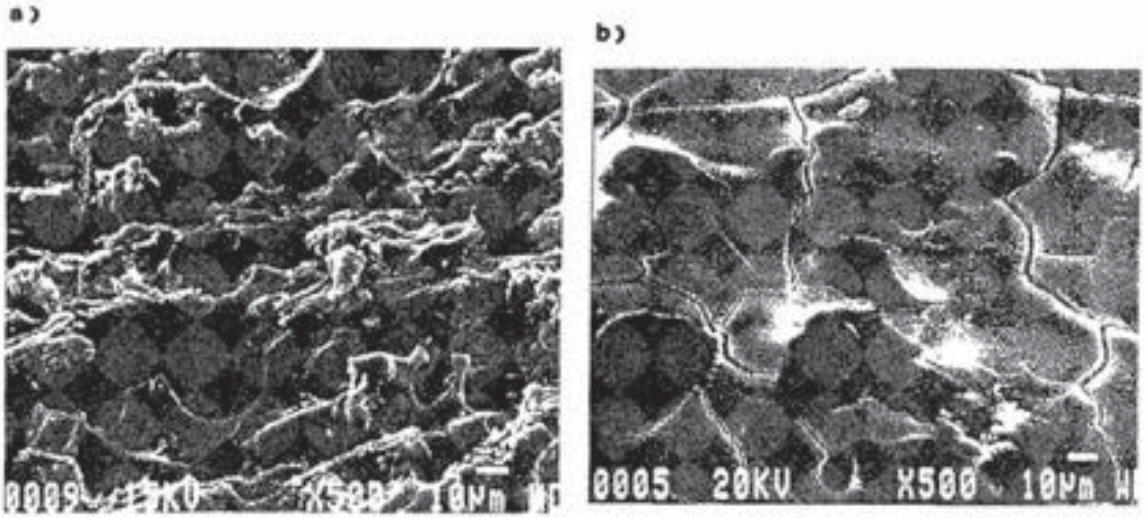


Fig. 8-59 (a) Polishing pad cross section. (b) Effect of pad conditioning.[47] Reprinted with permission of Solid State Technology. Published by PennWell.

pad surface of a Rodel IC-1000/Suba IV stacked pad, with and without conditioning. The SEM of the non-conditioned pad shows that its pores are completely clogged.[47]

The simplest form of pad-conditioning is done with a nylon-toothed brush. This is adequate for large particle removal or for conditioning soft, woven (or high-nap) pads. However, for the harder, molded closed-cell polyurethane pads (i.e., the Rodel IC-1000 or IC-1400 series), pad-conditioning is carried out by sweeping the pad surface with a disc containing 80-120-grit size diamonds embedded in a nickel plate (Fig. 8-60). Diamond, the hardest material available, is well suited for conditioning the hard-polyurethane-pad materials, but it is also brittle and can fracture. Alternative conditioning-abrasives have been pursued, since diamond particles may fracture and break away from the conditioning disk and become embedded in the pad. Since 80-grit diamonds are 100-200-microns in diameter, this can lead to severe scratching of the surfaces being polished.

Several approaches are used to mitigate the scratching problem. First, superior types of bonding mechanisms have been developed, which chemically bond the diamonds to the base of the conditioning-pad. Diamonds adhere better to such pads than to those that employ a plating technique for bonding. Second, periodic cleaning of the conditioning wheel is also carried out. Finally, some pads (such as the Rodel IC-1000) are fabricated with grooves created in the top surface of the pad. These grooves not only provide the pad with more slurry-carrying capacity, but dislodged 80-grit diamonds tend to be collected in the deep surface

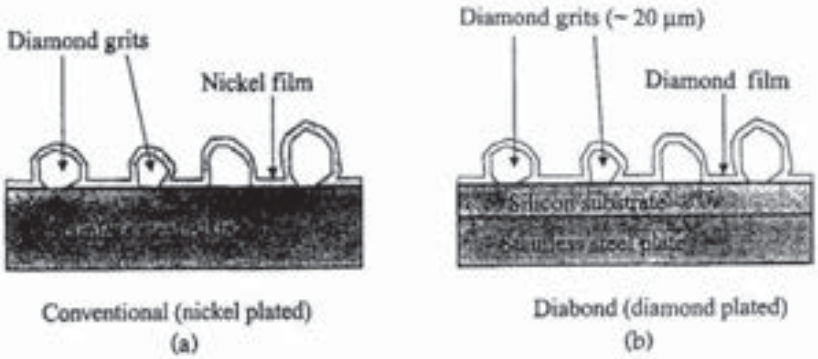


Fig. 8-60 Surface of pad conditioner showing bonded diamond grit.

grooves. Since the ridge tops of the pad between the grooves perform the polishing work (and not the pad surface within the grooves), the collection of the diamonds in the grooves minimizes scratches on the wafer surface. The pad is also cleaned with a high-pressure spray between wafers to dislodge and wash away the diamonds embedded in the grooves. (This high-pressure spray is more critical near the end of a pad's life, because at that point the grooves have become shallow. At that point, the diameter of the diamonds collected in the shallow grooves can exceed the groove depth and scratch the wafers). By using all of these techniques, diamond loss (and resultant microscratching), now represents much less of a problem.

Many polishers incorporate a continuous *in-situ* pad conditioning capability (a unit that moves across the pad during polishing, see Fig. 8-61 and Fig. 8-62).[48] This allows each wafer on the platen exposure to a pad with the same amount of conditioning, resulting in a more uniform removal-rate from wafer-to-wafer.

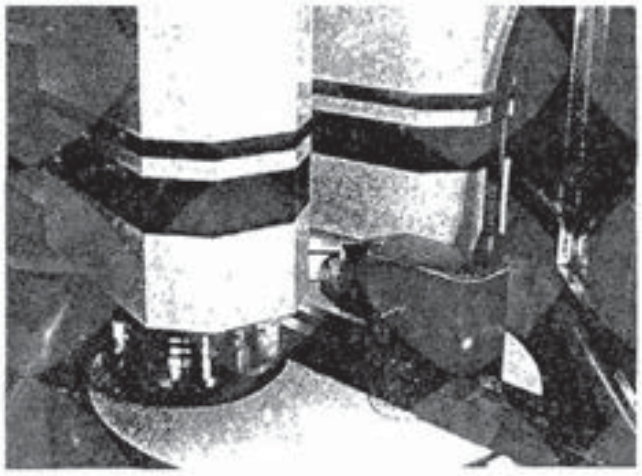


Fig. 8-61 In situ pad conditioner. Courtesy of Strasbaugh Inc.

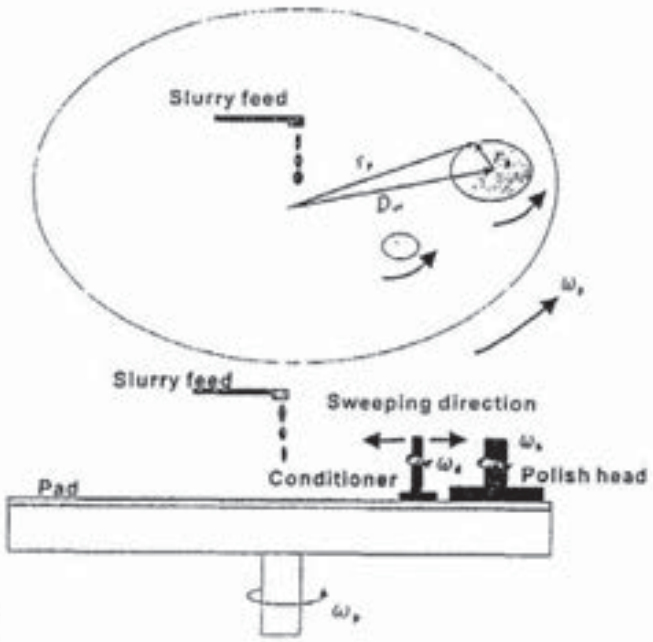


Fig. 8-62 *In situ* pad conditioner.

New pads also require a break-in conditioning-cycle to establish a constant removal-rate. In this procedure a set number of dummy wafers are polished for the purpose of stabilizing the pad condition and performance. Break-in conditioning-cycles can consume a significant fraction of the pad's useful life, and the polishing tool is not available for productive use while this is in-progress.

As a final note, it should be pointed out that conditioning of the pad is used to reduce the variation in removal-rate that occurs over a pad's lifetime. Yet the act of pad conditioning becomes yet another variable. As a result, elimination of pad conditioning would be an approach toward improving process stability. Use of a fixed-abrasive, one-use-pad, that is disposable or continuously replaced, is a potential path toward that goal.

8.6.8 Slurry-Free Pads

Slurry-free (or *fixed abrasive, FA*) CMP pads have also been developed. In these pads the traditional slurry-and-pad consumable-set is replaced with a pad in which the abrasive is attached to the pad. The main benefits of slurry-free pads are higher degrees of planarization and lower cost-of-ownership. The reduction in the cost-of-ownership is due to a number of factors. First, there is the reduced cost of consumables. Second, the cost of slurry handling is eliminated. Third, there is no longer a need for a slurry-distribution system. Finally, the problems of slurry-stability and solution-variation are eliminated. To date, most development work on slurry-free pads has centered on ILD and STI applications, but programs for slurry-free copper-CMP are also underway.

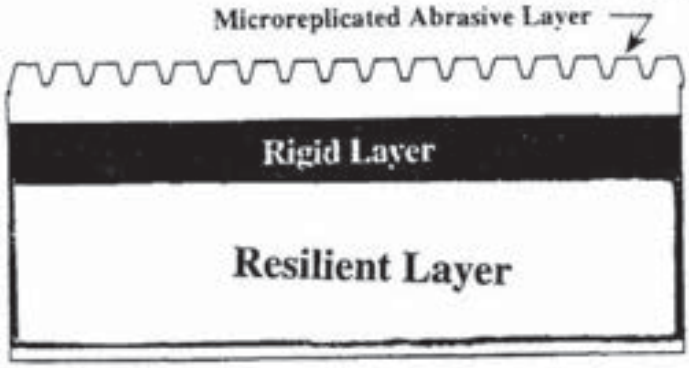


Fig. 8-63 Slurry-free (Fixed abrasive) pad structure.

A pad supplied by the 3M Co. embeds the abrasive material in the pad itself, in the form of a precise micro-replicated abrasive surface and a bi-component sub-pad system (Fig. 8-63). The abrasive surface can be constructed in one of several configurations chosen for specific performance criteria. Figure 8-64 shows two of the surface features currently in use.[49] For dielectric applications the slurry-free pad is used with a simple basic solution. For metal-CMP, a

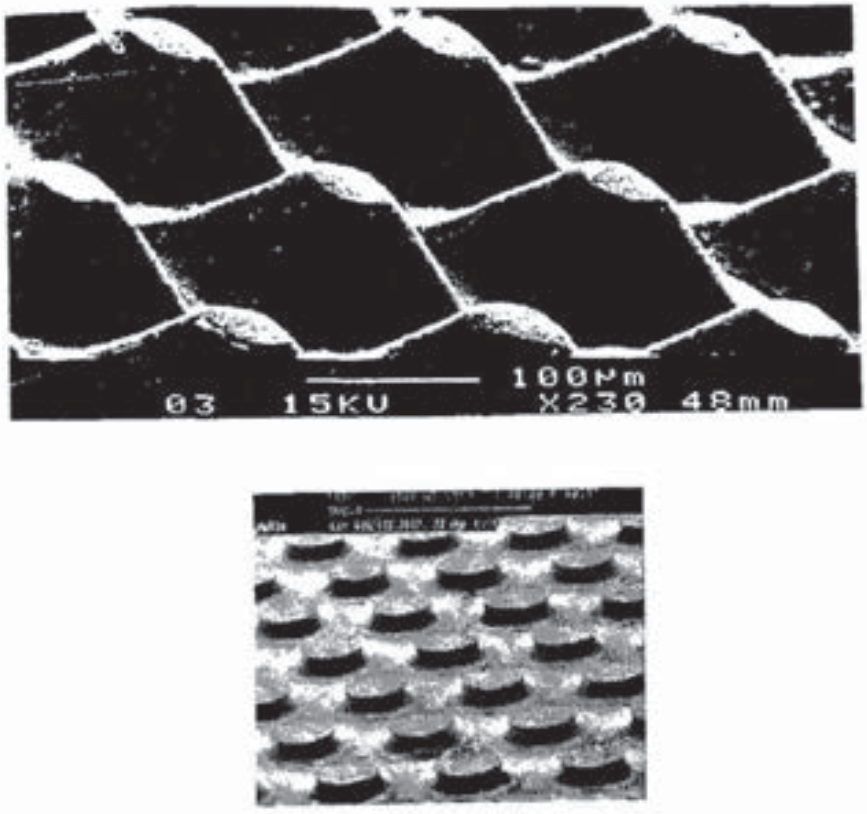


Fig. 8-64 Examples of slurry-free pad surface structures. Courtesy of 3M Co.

a specially formulated oxidative solution is needed, with a pH value of 2 to 7.

These pads are composed of four layers: a micro-replicated abrasive; a rigid layer; a resilient layer; and a self-adhesive backing. These slurry-free pads can be used either on a moving-platen (rotary or linear) or a fixed-platen. In early 2000 Applied Materials and 3M entered a strategic agreement to jointly develop a slurry-free pad for the Applied Mirra® multi-platen rotary CMP-tool. 3M also has a marketing alliance with pad supplier Rodel. SpeedFam-IPEC also announced an effort to develop a slurry-free pad for their rotary/orbital CMP tools. There is some concern about the consistency of slurry-free pads for rotary CMP tools throughout the pad life. As the pad wears, the surface area tends to change, which may cause a non-stable process.

These concerns are avoided in web-format pads for fixed table, roll-to-roll pads. The use of a slurry-free pad in a fixed-table (platen) CMP system was pioneered by Obsidian Corp. In Obsidian's web-based tool, a roll of abrasive pad material is spooled across a platen and incremented after each wafer. No slurry or pad-conditioning is required.[50] (Fig. 8-65). However, one concern about this slurry-free approach is the uniformity of polishing performance from roll-to-roll.

Slurry-free pads have demonstrated excellent planarization capability. This capability allowed a slurry-free STI CMP-process performed on an Obsidian 501 CMP tool to demonstrate exceptionally tight nitride ranges and minimal dishing (Fig. 8-66).[51] That is, the removal of nitride during the trench oxide CMP process is very uniform across the wafer, with only a 250-Å variation in the remaining nitride films after the polish step. The dishing in the field oxide regions was also very small, with the maximum dishing being less than 400-Å.

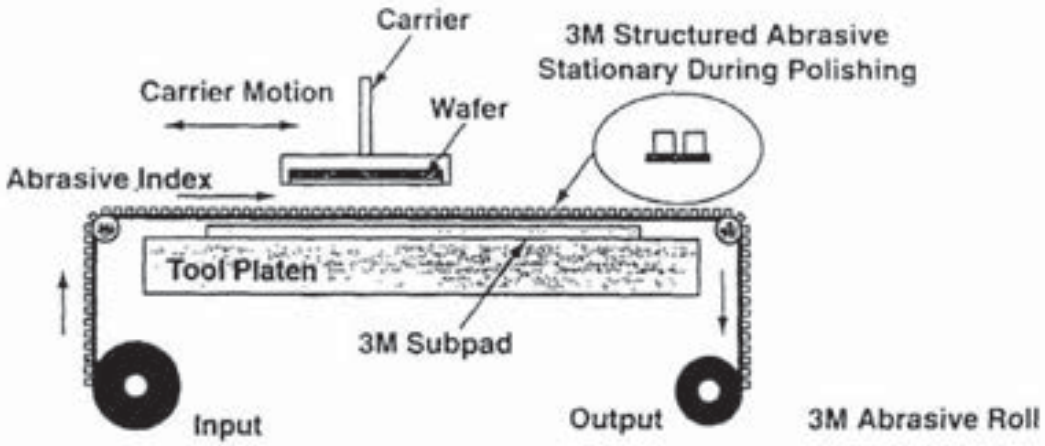


Fig. 8-65 Schematic of linear web format CMP tool, with indexing, slurry-free pad.

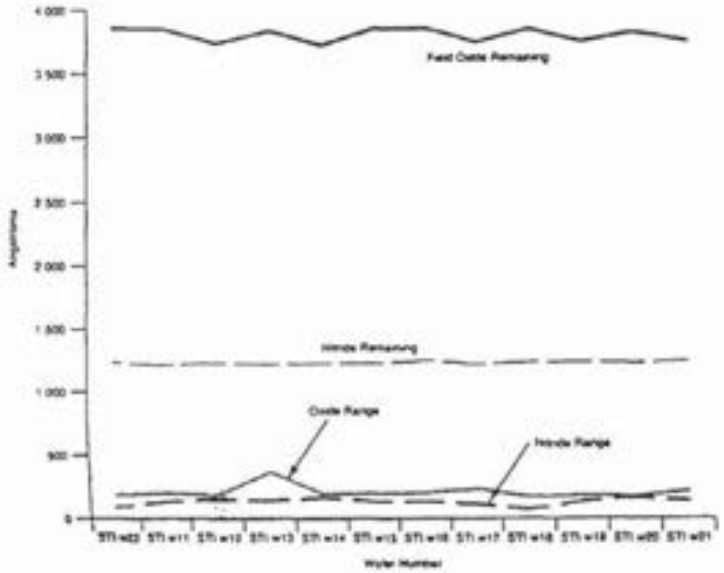


Fig. 8-66 Fixed abrasive polishing results on Linear CMP tool (roll-to-roll pad). Test performed on the Obsidian 501 CMP tool at the 3M Applications Lab. Courtesy 3M Co.

8.6.9 Endpoint-Detection in CMP

Since CMP is a process for reducing thickness at selected locations on a wafer, it is necessary to be able to determine when the process has reached its *endpoint*. There are two types of processes that need to be controlled: **1**) polishing a dielectric to a known final film thickness; and **2**) polishing an inlaid, usually multi-layered, metal film and stopping on the surrounding dielectric (after all of intra-inlaid metal is removed). The control of the first of these processes, that of polishing an oxide, can only be accomplished by making measurements on the film to be polished before the CMP process, and again after the wafer has been polished for some pre-determined time (based on the pre-polish film measurements). This is termed an *in-line endpoint-measuring* process. Another approach is to measure the film thickness as the polishing process is proceeding. This is called a *real-time, in-situ endpoint-detection* process. At the present time it is possible to control the second type of CMP process listed above (but not the first) by using *in situ endpoint-detection.* Nevertheless, both techniques will be described.

In the *in-line endpoint-measuring* method used for controlling dielectric-CMP processes, a pre-CMP measurement of the dielectric film to be polished is performed.[82] This measurement is made over an area containing dense device features to determine the thickness, T_o, shown in Fig. 8-17. Based on the

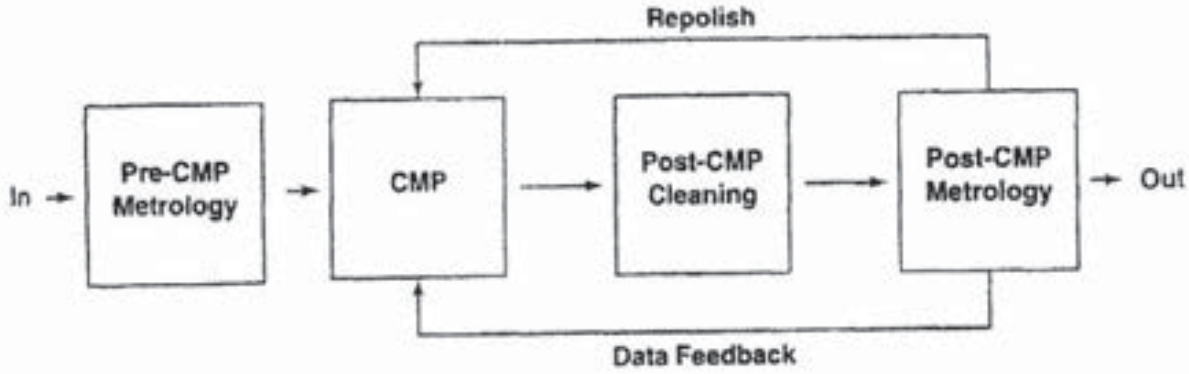


Fig. 8-67 *In-line* endpoint detection method used in CMP (i.e., polishing, followed by measurement to confirm thickness - and re-polishing if necessary - is the cycle used with *in-line* endpoint detection).

measured value of T_o, the polishing time (t_{final}) needed to reduce the film to a final required thickness, T_{final} (see Fig. 8-17) is calculated. The value of t_{final} is chosen from CMP calibration data based on applied pressure, rotation speed, and average pad-life-degradation. This produces a value of T_{final} centered within the allowed process window. Various factors may alter the actual value of T_{final} realized at t_{final} (e.g., by as much as 20%, 1-σ, during process development). If the material removed by t_{final} exceeds the allowed limit, the wafer must either be scrapped or more dielectric material must be deposited. If the remaining thickness is excessive, the wafer can be returned to the polisher for more removal. In fact, a typical manufacturing approach to oxide-CMP in which in-line endpoint-detection is employed, is to deliberately *underpolish* the SiO_2 layer. The film is then measured and polished a second time, using an adjusted polishing time. The cycle is repeated until the measured thickness is within specification limits. This procedure is often carried out on a look-ahead, or "pilot" wafer for each lot (Fig. 8-67), and the remainder of the lot is polished according to the time calculated when polishing the pilot wafer.[53]

This loop is the basic approach used in dielectric CMP (Fig. 8-68), but it has some drawbacks. The most significant are that the technique is time-consuming, measurement-intensive, and reduces the overall time the polisher can be kept in use. In addition, there is a risk of wafer breakage when being removed for measurement. Some thickness metrology tools are designed to be placed inside the polishing tool housing as another subsystem (e.g., the ITM® tool of Nova

Measuring Instruments). A central handling mechanism transports the wafer to the measurement position and then to a scrubber, then back to polish or to the output cassette. Multiple points on the wafer can be measured with a small optical head. This head (rather than the wafer stage) is moved, providing about 5 times faster movement.

The use of a *real-time, in situ endpoint-detection* method allows the problems of in-line endpoint-detection to be circumvented. *In situ* endpoint-detection is a way of identifying the completion of a polish step without having to move the wafer outside of the tool platform. *In-situ* endpoint-detection would permit the user to stop at the desired remaining thickness without having to re-polish. This would also improve throughput and increase the process window. Several methods have been developed to perform *in-situ* endpoint detection in CMP, including: **1**) front-side laser interferometry or spectrometry; **2**) optical-sensing of multiple-points on the backside of the wafer; and **3**) monitoring the current

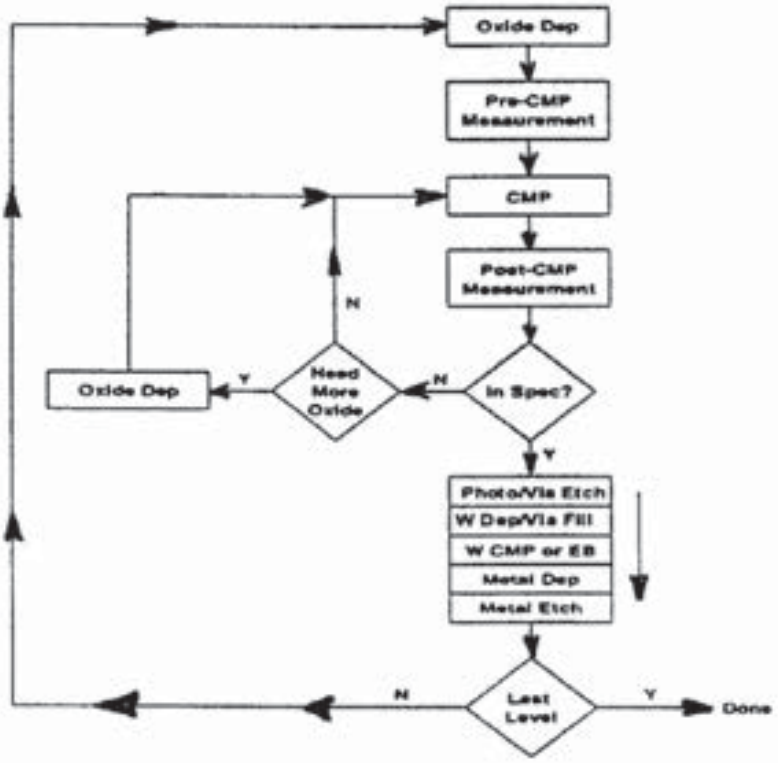


Fig. 8-68 Implementing the *in-line* endpoint-detection method used in CMP.

supplied to the motor of the wafer carrier. Unfortunately, as noted above, no satisfactory *in situ* methods yet exist for polishing homogeneous dielectric films to a known final thickness.

The *wafer-frontside laser-interferometry technique* is used by Applied Materials in their Mirra® tool, the *spectrometer reflectivity technique* on the Lam Research Teres® tool and in the Speedfam Auriga® tool. The name of the Applied Materials endpoint system is the *in-situ removal monitor* (ISRM).[54,55] Each polishing platen has a red laser and a detector mounted on the polishing table. A window in the polishing pad allows the laser to be aimed at the top surface of the wafer at 45° to the wafer surface. A detector is used to sense the reflected light (also collected at 45° from the surface). As the laser/detector passes under the wafer, the laser light passes through the clear window in the polishing pad, and reflections are sampled. The intensity of the reflected light is measured as a function of time (the light-intensity vs. time curve in Fig. 8-69).[56]

A typical reflectance curve for polishing Cu/Ta films on an oxide dielectric is shown in Fig. 8-70. After the maximum reflectance is reached, the process starts to break through the Cu layer to the semi-transparent Ta layer. At that point a sharp drop in reflectance occurs. Eventually, the copper filling the trenches continues to reflect a small constant amount of light.

The Applied Materials Mirra® CMP tool uses sophisticated signal processing/ detection algorithms to reduce noise from the slurry and pad surface condition (Fig. 8-71). In development tests, endpoint accuracies to within 20-nm were reportedly achieved. In W-CMP polishing, the ISRM system measures the reflectivity of the metal film and signals the transition from metal film to oxide. An over-polish time is specified to ensure complete removal of the liner film.

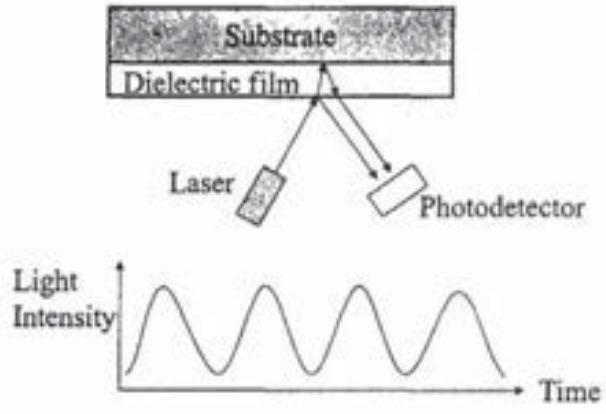


Fig. 8-69 *Wafer-frontside laser-interferometry* endpoint-detection for dielectric CMP.

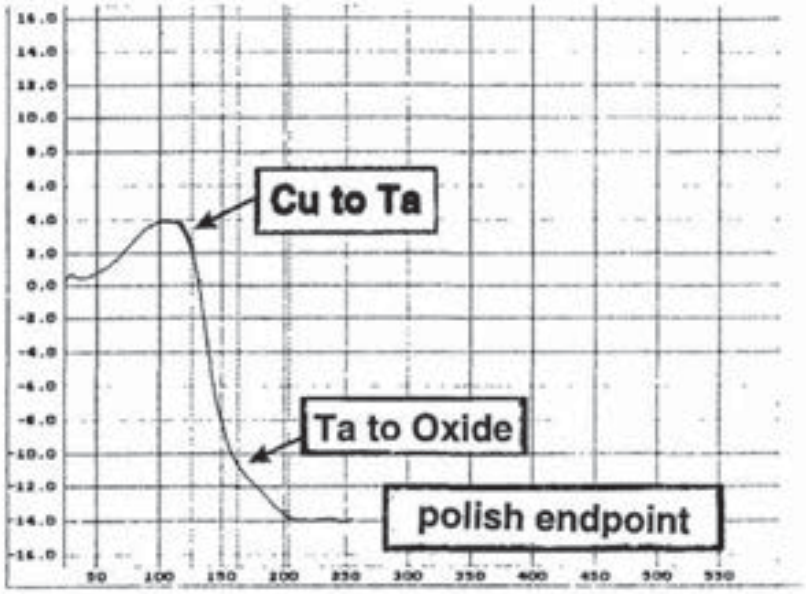


Fig. 8-70 Typical reflectance curve for polishing a Cu/Ta/oxide film.

In the Lam Research Teres® tool, a *multiple-wavelength, spectrometry-based endpoint-detection system* is employed. It can continuously monitor the thickness of transparent films as they are being polished. It can also measure the thickness of very-thin metal films. Thus, during a Cu CMP-process, the endpoint of the Cu polishing step can be sighted before all the Cu is removed. This allows a set of process conditions that rapidly removes Cu to be initially used. When the point of complete Cu removal is detected, the polish conditions can be switched to those that result in a slower Cu removal rate. This allows the Cu to be completely

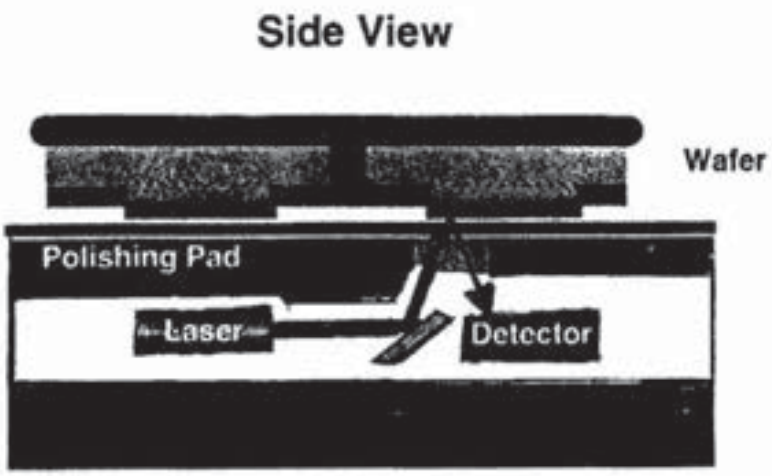


Fig. 8-71 Schematic of the Mirra® laser-reflectometer light path. Courtesy Applied Materials.

removed over the underlying Ta liner film without significantly eroding the exposed Cu in the trenches. In the Speedfam Auriga® multi-head polisher, a halogen lamp is used as the light source and diode arrays are employed to detect the light reflected from the surface of the wafer being polished. Endpoint in W-CMP is determined from an intensity change in the reflected-light as the exposed surface changes from tungsten to oxide.

The *optical monitoring of the wafer backside* as an endpoint-detection technique for monitoring the oxide-CMP process has been developed by Luxtron.[57] In this approach, the wafer's backside is illuminated with infrared radiation (to which silicon is substantially transparent). The radiation is therefore able to sense the film thickness on the wafer's front side as it is being polished.[58] By mounting several sensors onto the wafer carrier, multiple points on the wafer can be monitored. The film thickness is measured using the basic principles of optical interferometry as was described in earlier paragraphs. For wafers with no metal on the front surface, this technique is quite straightforward. However, if there is some metal on the front surface (e.g., in a multilevel-metal process), the situation becomes more complicated, because the metal will block some of the incident light. Fortunately, there is generally sufficient space between device features to allow light to pass. Figure 8-72 shows a schematic of a single sensor probe. Two optical fibers connected to an infrared emitter and detector are mounted in a modified wafer-carrier of a commercial CMP tool. Electrical communication from the sensors is accomplished via a wireless link. As many as four sensors can be mounted in a single carrier.

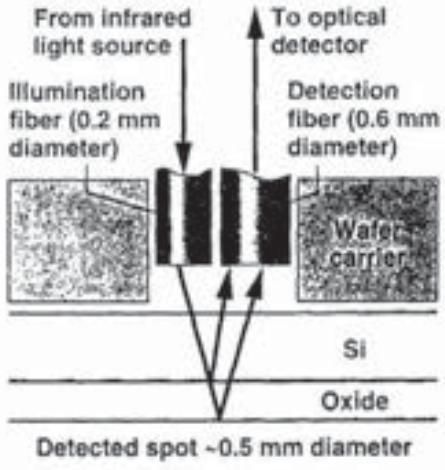


Fig. 8-72 Schematic of an optical sensor embedded in a wafer carrier.[57] Wafer is backlighted with IR light to measure films on the front surface. Reprinted with permission of Semiconductor International.

The *motor-current-monitoring technique*[57] is a production-proven endpoint-detection method that works well when polishing a material down to a stop layer (such as the polishing of a CVD-SiO_2 film on a silicon nitride stop layer in shallow-trench isolation,[58] and the polishing of W over oxide in W-plug CMP processes[57]). It is also relatively easy to implement. Circuitry such as a *current shunt* or a *Hall-effect probe* is used to monitor the current supplied to the motor that rotates the wafer-carrier (Fig. 8-73a). Since the wafer-carrier is driven at a constant rotational speed to maintain a constant polishing rate, the drive-current is varied to compensate for any load changes on the motor. This makes the motor-current sensitive to frictional changes at the wafer surface. The largest changes occur when one material has been polished away, exposing a layer which has different polishing characteristics (e.g., that could function as a CMP-stop layer, see also Fig. 8-73b). Since there is no layer breakthrough in oxide-CMP (as can be seen in Fig. 8-25a), the motor-current-monitoring technique is not appropriate for such applications. Luxtron and Aplex are vendors of such motor-current endpoint controllers (e.g., the Luxtron Xinix Model 2450®).

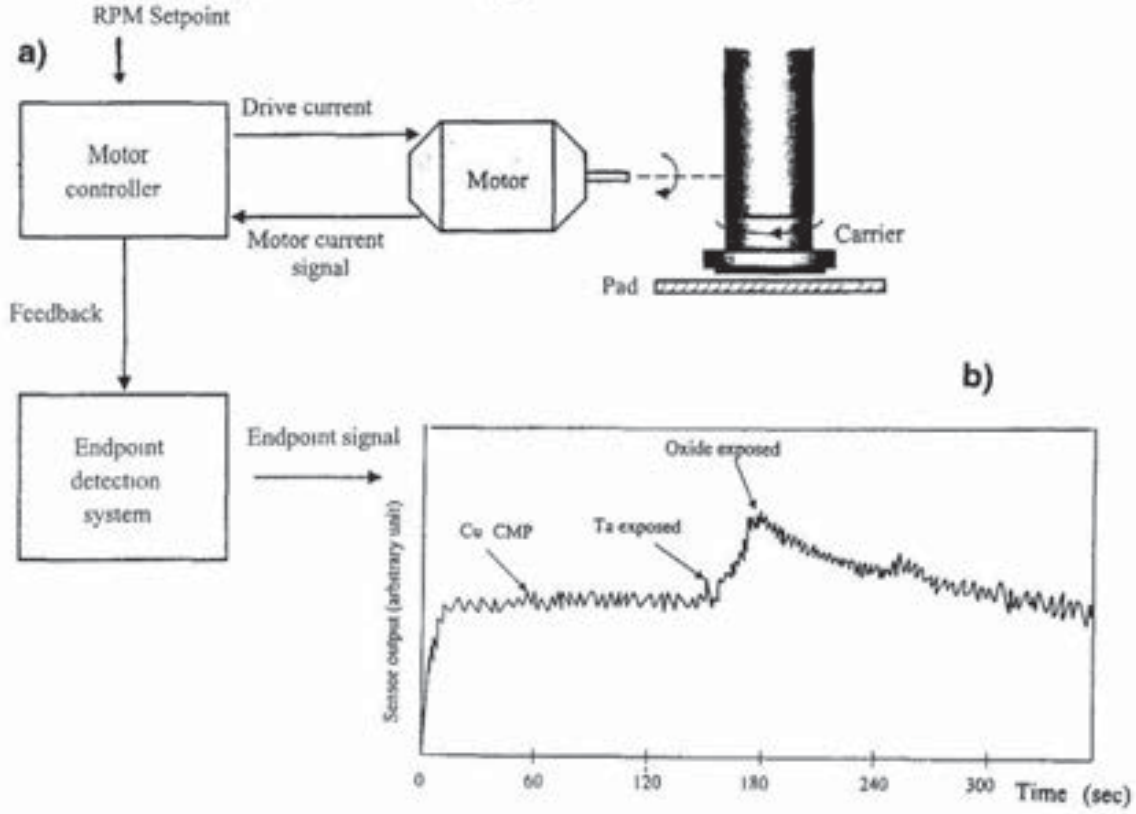


Fig. 8-73 a) Motor current endpoint detection. b) Motor current output during Cu/Ta/oxide CMP process. Courtesy of Aplex.

8.7 CLEANING ISSUES IN CMP

CMP is an inherently dirty process that can introduce yield-reducing defects to the wafer. That is, after a polishing step, the wafer-surface contains slurry particles, abraded pad-material, and debris that was removed from the wafer-surface. Residual slurry specifically introduces foreign particles, metal contaminants, and chemicals; all of which are undesirable for the remaining steps of chip fabrication. If the slurry particles alone were allowed to remain on the wafer, such contamination could obstruct electrical connections, create voids during subsequent oxide- or metal-deposition steps, or cause corrosion of exposed metal areas. Some early resistance to the adoption of CMP stemmed from such contamination concerns. As a result, significant effort has been expended to develop post-polish cleaning techniques so wafers returned to the cleanroom after CMP are not contaminated.[60]

To prevent CMP-based contamination, wafers are subjected to post-CMP cleaning steps that return the wafer-surface to an acceptable level of cleanliness. While the main objective is to provide a wafer-surface that is free of slurry particles, the cleaning process must also remove organic-residues and trace metal ions. The cleaning process must do this without introducing further defects. In the case where inlaid metal is exposed (e.g., in a W-plug CMP process or a in Cu-damascene CMP process), corrosion must also be controlled or avoided.

Since CMP is a wet process, CMP cleaning-technologies are mostly aqueous. The post-CMP wafers are kept wet as they are moved from the polishing tool to the cleaning station. The initial wafer-rinsing (after the polishing is completed), is done while the wafers are still on the polishing pad. Spray rinsing is also done as soon as the wafers are lifted from the pad. If wafers are exposed to air for even one or two minutes before such spraying (e.g., while waiting to be unloaded), subsequent cleaning becomes difficult (and sometimes impossible). Immediate removal of metal-polish slurries is also critical, because these can continue to react and remove material even after polishing has ceased.

Current models of CMP polishing tools either have an on-board (i.e., fully integrated) cleaning station, or they have the capability to interface with a stand-alone cleaner (in the form of a cassette-to-cassette track with stations for brush scrubbing [one or more], rinsing, and spin-drying). If a stand-alone cleaner is employed (as was common in early CMP processes), the wafers are transferred from the polisher to the cleaner by a human worker. If the cleaner is integrated with the polisher, (the most common method in the latest generation of CMP

tools), automated wafer transfer is used (i.e., the wafer is moved from the polisher to the cleaner by a robot).

To prevent additional wafer and slurry contamination, the number of airborne particles should be minimized by filtering the air fed to the CMP tool with the same HEPA filtering techniques used for the cleanroom environment. The cleanliness of the DI water used to rinse or submerse the wafers during cleaning is also an issue. Tubs or baths in which wafers are kept wet must be continuously refreshed with new DI water. The CMP tool should automatically rinse all surfaces that come in contact with slurry, including the polishing table, drain tubs, carriers, and pad conditioner. These areas should also be cleaned periodically to prevent dried slurry buildup. As mentioned previously, slurry is composed of very small abrasive particles, and filtering by conventional methods is not easy. Nevertheless, such filtering is needed because large agglomerates can develop within the slurry and have serious device-defect consequences.

8.7.1 Mechanisms of Particle Adhesion and Removal from Surfaces

Particles are deposited on wafers as aerosols from the air or as particles present in the liquids in which wafers are immersed. A description of the techniques used to remove them (e.g., after a CMP polishing step) rests on an understanding of how such deposited particles adhere to wafer surfaces. Thus, a short description of this phenomenon follows.

If a solid particle wanders close enough to the vicinity of a solid surface, the *Van-der-Waals force* (*VdW*) can attract and capture it. The particle is then *physisorbed* on the surface. The Van-der-Waals force is universal, and is thus effective in trapping all types of solid particles (and gas molecules, as well). Other forces that physisorb particles include *static charge on the particle*, and *capillary adhesion* (the force exerted by a liquid film between the particle and the surface), but the VdW force is usually the strongest of these. In fact, it can be so strong that the particle may elastically (and sometimes even plastically) deform so that the area of contact between the particle and the surface is increased. This makes such binding even stronger. Nevertheless, particles held by VdW forces can be removed by the techniques described here.

There are three methods used to remove physisorbed particles: **1**) *electrical repulsion of the particle and the surface*; **2**) *chemically assisted removal* (particle dissolution, particle oxidation [and then dissolution], and lift-off by slight etching of the surface); and **3**) *mechanical dislodgement* (sliding and rotating the particle

by the application of a shear force). These mechanisms are described in more detail in the next sections.

Note that cleaning by immersion-only is not effective for post-CMP cleaning, due to the particle adhesion forces and boundary layer effects. That is, merely rinsing in DI-water or immersing the wafers in a wet-chemical solution has not been found to adequately remove slurry particles from the wafer-surface.

8.7.1.1 Electrical Repulsion of the Particle and the Wafer Surface: Besides the particle adhesion forces discussed above, other inter-particle (or particle-to-surface) forces exist which can be either attractive or repulsive. The most important of these forces are those involving the charge near the particle surface. That is, the outer-most charge near the surface of a particle most influences the way a particle will interact with other particles or surfaces. Particles and surfaces immersed in a liquid can be charged by the adsorption of ions from the solution. The charge thus adsorbed on the surface is balanced by an equal but oppositely charged layer in the adjacent liquid, resulting a so-called *electric double layer*. (Fig. 8-74) The potential associated with this adsorbed charge on a particle or surface is termed the *Zeta potential* (ZP) and is also depicted in Fig. 8-74. Since the particle or surface adsorbs ions from the liquid, the ZP of the particle is also

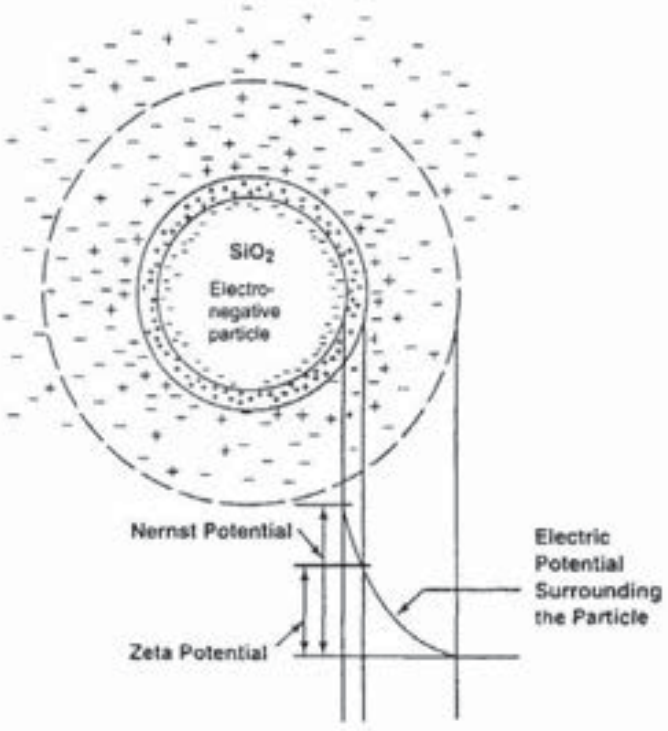


Fig. 8-74 Slurry chemistry electrostatic layers surrounding an SiO_2 particle (including a depiction of the electric dipole layer).

consequently dependent on the pH of this surrounding media. For example, Fig. 8-75 shows the ZP of a variety of particle types (SiO_2, Al_2O_3, W, poly-vinyl alcohol [PVA], and Si_3N_4), as a function of the pH of the liquid in which they are immersed. The silica particles have a negative potential of –30-mV if they are immersed in a liquid with neutral pH (i.e., pH = 7). Also note that their ZP (as well as those of the other particles shown in Fig. 8-75) becomes strongly negative as the pH is increased above pH = 7. (The ZP is also affected by the addition of a surfactant to a liquid, and thus the ZP of a solid-surface can likewise be manipulated by the addition of surfactant to the liquid.)

To prevent slurry particles from sticking together, it is best to immerse them in a solution that will stabilize their respective ZPs to either a positive or a negative value. This will ensure that the particles will repel each other (i.e., this repulsion - see Fig. 8-76 - will counteract such attractive forces as the Van-der-Waals forces). Consider the case of silica particles in a silica-based slurry. Since silica particles tend to have negative values of ZP, the net-negative electric charge near the particle-surface produces a force that tends to make the silica particles repel each other. However, to ensure that these repulsive forces are strong enough to overcome the opposing attractive forces (such as VdW), a stabilizing chemical that increases the pH of an aqueous solution (e.g., NH_4OH) is added to the slurry.

From the data shown in Fig. 8-75 it can be seen that to keep the silica particles in a slurry from agglomerating, they should be immersed in a liquid with pH values > 7 (basic). On the other hand, for slurries containing alumina particles, one might want to consider using a slurry solution with a low pH (i.e., < 7 [acidic]), because in such an acidic solution the alumina particles exhibit a positive ZP.*

To remove particles from a surface, consider the case of removing silica-slurry particles from a wafer-surface covered with oxide (i.e., a condition that exists at the completion of an oxide-CMP step). Since the ZP of the SiO_2 surface is also negative (at neutral pH), it is preferable to use a cleaning chemistry with a high pH (i.e., > 7). This causes the silica slurry particles adhering to the oxide

* We noted earlier that the CMP process also consumes large volumes of DI water. Work has thus been carried out to develop methods of treating the CMP wastewater so that it can be re-used. One phase of such reclaim treatments involves removing the slurry particles from the wastewater. For the case of W-CMP waste water, it was found that by a simple adjustment of the pH of this water, the slurry particles would agglomerate into large clumps, making their capture and removal quite easy.

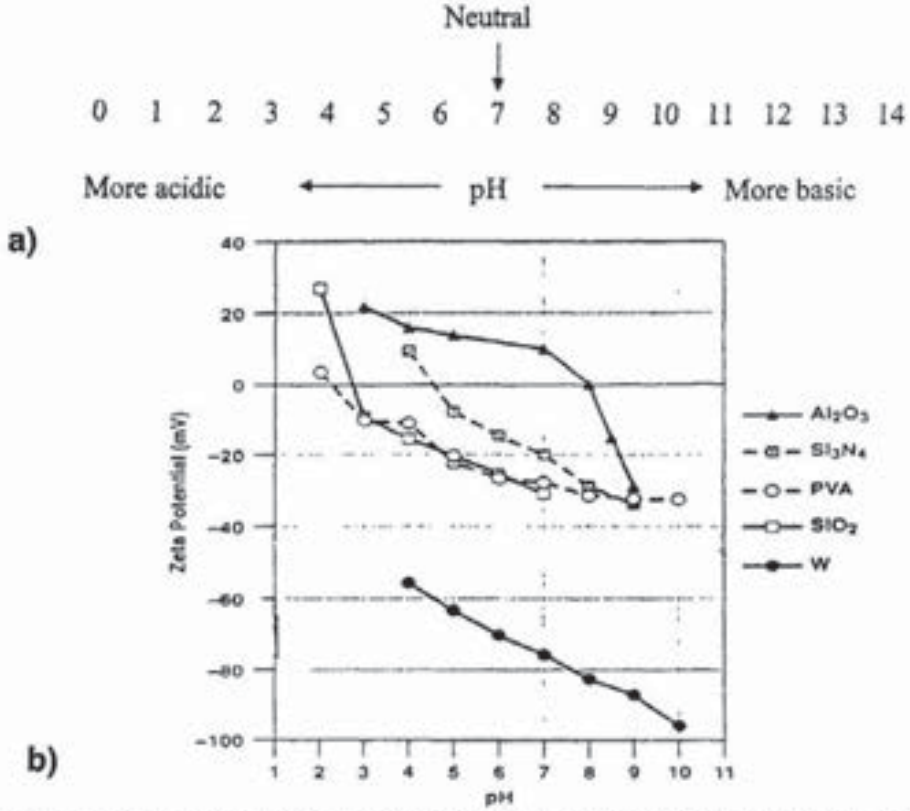


Fig. 8-75 a) Relationship of pH and acidity and alkalinity. b) Zeta potentials versus pH for particles of various types of slurry particles and materials in an aqueous solution.

surface to acquire additional negative charge, which increases the repulsive component of the force between particles and the surface, making any par-ticles adhering to the surface easier to remove. Note that in most cleaning solutions, chemistries are extremely dilute (i.e., 1:50, chemical:DI water).

It should also be mentioned that for many CMP processes, including W-CMP and Cu-CMP, the exposed wafer-surface at the conclusion of the polish-step contains more than one type of material (i.e., not just oxide). For example, when the CMP-process for polishing W-plugs (or Cu-interconnect-lines) is completed, more than 50% of the exposed wafer-surface area consists of oxide, and the remainder consists of metal. Nevertheless, since the majority of the exposed surface is still oxide, the most effective cleaning chemistries will still be those that remove slurry particles from oxide-surfaces. Hence, cleaning solutions used to remove silica-slurry- particles from these surfaces should have pH values > 8.

The same considerations apply to the cleaning of slurries that use alumina particles. That is, as shown in Fig. 8-75, alumina particles have a positive ZP if they are immersed in a solution with a low (acidic) pH (i.e., < 8). Thus, if an acidic cleaning solution were to be used, the alumina particles would be attracted to the electronegative oxide surface. To prevent this situation, the cleaning

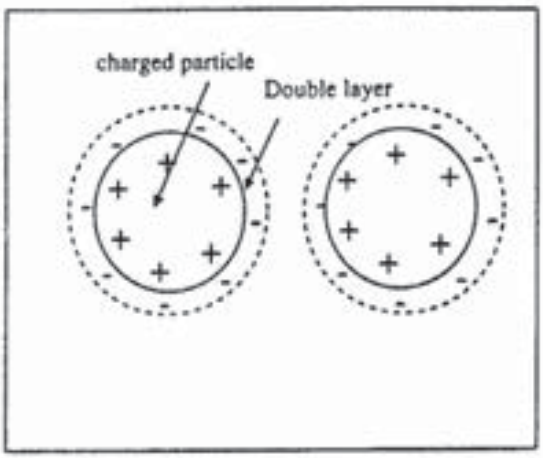


Fig. 8-76 Electrostatic repulsion keeps particles separated by repulsion of charged double layers.

solution must again have a high pH (> 8). This makes the ZP of the alumina particles negative, so they are repelled from the oxide surface on the wafer.

Figure 8-77a compares brush-cleaning results after oxide-CMP with and without NH_4OH in the cleaning solution. Figure 8-77b compares the results after W-CMP with and without NH_4OH. In both cases, the use of NH_4OH (which raises the pH of the solution) is seen to reduce defects by as much as 75%.

8.7.1.2 Chemically Assisted Particle Removal: The *chemically assisted* particle removal techniques are implemented in some wet-cleaning procedures (Fig. 8-78). For example, in the SC-1 step of the RCA clean (see Vol. 1, 2nd Ed. Chap. 5), several chemically-assisted particle removal mechanisms are believed to work in tandem. That is, the SC-1 solution removes some types of particles by dissolution (i.e., organic particles). For the case of cleaning bare silicon wafers, it removes other types of particles by slightly simultaneously oxidizing and etching the Si surface. That is, it undercuts the oxide under a particle, which reduces the contact area between the particle and the surface. This makes the particle more prone to being lifted off.

8.7.1.3 Mechanically Assisted Particle Removal: The *mechanical removal of particles* generally proceeds by momentum transfer. That is, shear forces impact a particle and dislodge it by either sliding it off the surface, or by rotating it until it detaches. As particle sizes decrease, they become more difficult to dislodge by mechanical force. Although the Van-der-Waals-adhesion force may get smaller, the electrostatic forces increase, and the efficiency of transferring translational momentum also rapidly declines. Mechanical particle removal techniques have traditionally involved: **1**) mechanical brush scrubbing; and **2**) vibrational

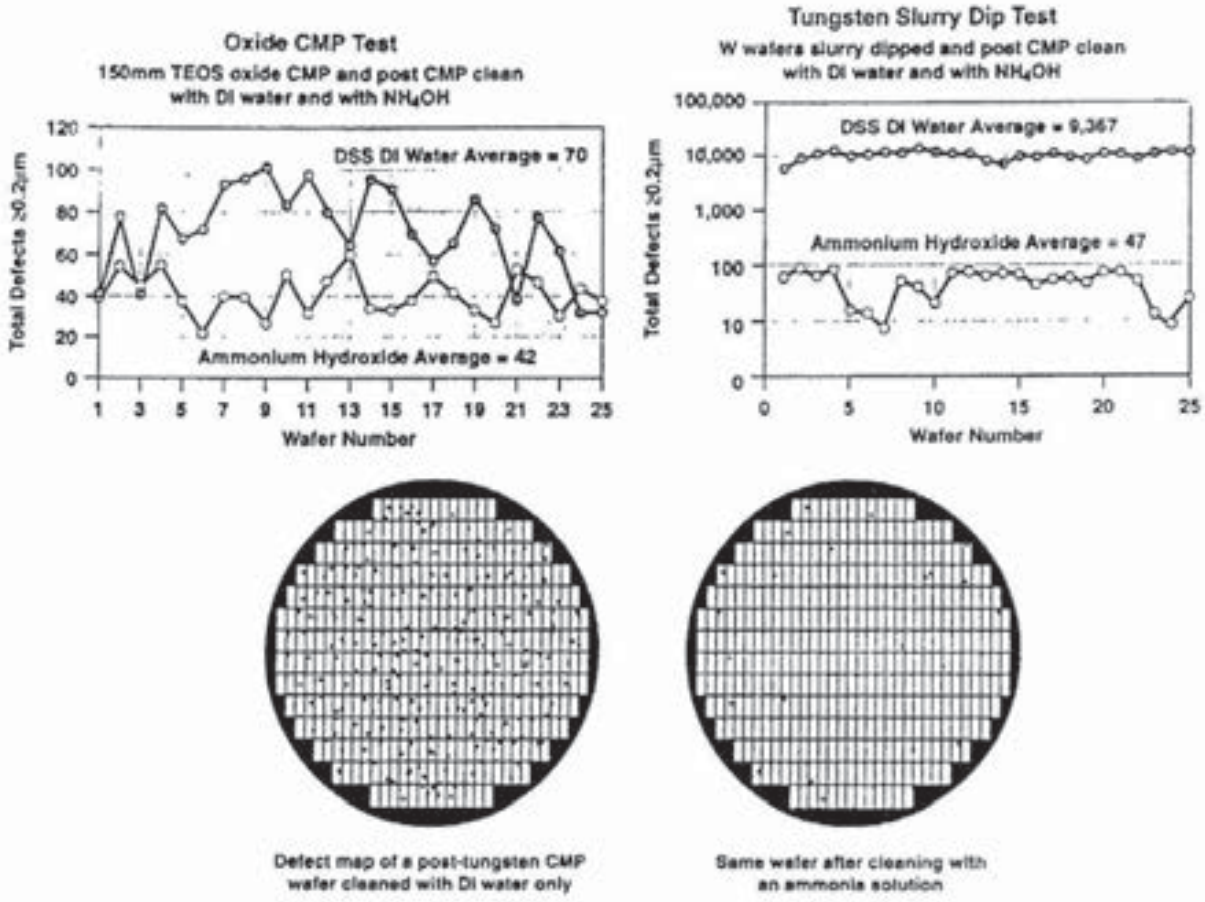


Fig. 8-77 Brush cleaning results comparing use of H_2O(DI) vs. NH_4OH: a) Oxide-CMP cleanup; b) W-CMP cleanup.

(*megasonic*) action. In fact, these particle removal mechanisms are employed in tandem in post-CMP cleaning processes. An explanation of how this is done will be provided in the upcoming sections.

8.7.1.4 Steric Stabilization for Suspending Particles in a Solution: Particles can be prevented from attaching to each other, or from settling on a surface, by the phenomenon of *steric stabilization.* That is, a molecular layer is attached to the particle surface (often a *dispersant* like a polymer or a surfactant is used for this role). This attached layer causes *entropic repulsion* as the particles approach each other (Fig. 8-79). This technique is a second method used to prevent settling and agglomeration in CMP slurries. If this mechanism is used to stabilize a slurry, care must be taken to ensure that the additive does not adversely affect any of the other slurry properties.[61]

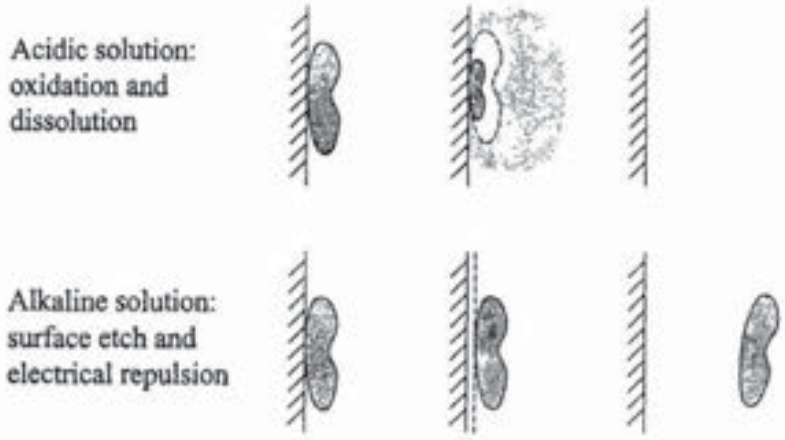


Fig. 8-78 Wet-chemical mechanisms that remove particles from surfaces: a) in a powerful oxidizing agent; b) in an alkaline solution. Courtesy of Semiconductor International.

8.7.2 Practical Particle Removal Processes: Brush Scrubbing

The most widely used method of post-CMP-cleaning involves brush scrubbing the wafers while they are sprayed with appropriate chemical solutions (or while they are immersed in these solutions). Brush cleaning systems have long been used in many wafer-cleaning applications. Despite its name, it actually uses a hydrodynamic drag to exert a removal force on the surface-particles (mechanical dislodgment). Brush bristles do not contact the particle or the surface, but rather act as oars or paddles that push liquid across the wafer surface.

While the brush scrubbing technology is not new, some modifications had to be made to allow a stand-alone brush scrubber to be integrated with a CMP

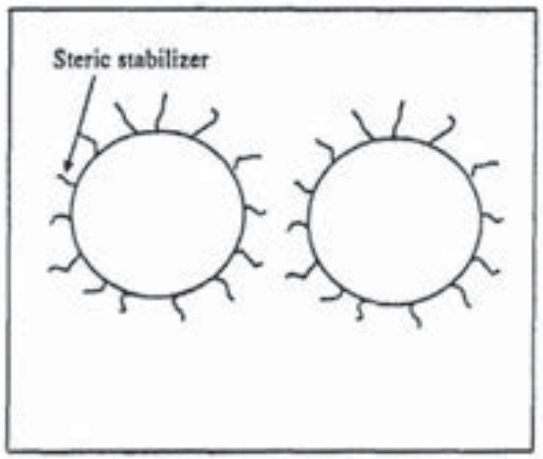


Fig. 8-79 Steric-stabilization of particles in aqueous solution. Particles are stabilized by attaching a molecule to the surface. The molecular layer causes an entropic repulsion as the particles approach each other.

polishing tool. For example, the current technology of choice for both oxide- and metal-CMP employs simultaneous double-sided brush scrubbing (see Fig. 8-80).[62] This simultaneously scrubs both the front and back of the wafer. Another change involves the fact that the scrubbing equipment must function reliably in the presence of both basic slurry chemistries (e.g., KOH and NH_4OH), and acidic cleaning chemistries (such as citric acid and HF), in order to optimally remove both metals and mobile ions. Alumina-bearing slurries used for tungsten-polishing more readily adhere to the wafer surface and are generally more difficult to remove. Particles as small as 0.12-μm have reportedly been effectively removed by brush scrubbing.

Particles are dislodged and they then stay suspended in the liquid until they are pushed over the edge of the wafer into the drain, as liquid is continuously delivered to the wafer surface. Dislodged particles, however, must be prevented from re-adhering to the wafer surface or the brush material. This is done by controlling the electrostatic forces between the particles and the wafer or brush surfaces. As described earlier, particle re-adhesion to both the wafer surface and the brush is suppressed by keeping the pH values of the cleaning liquid greater

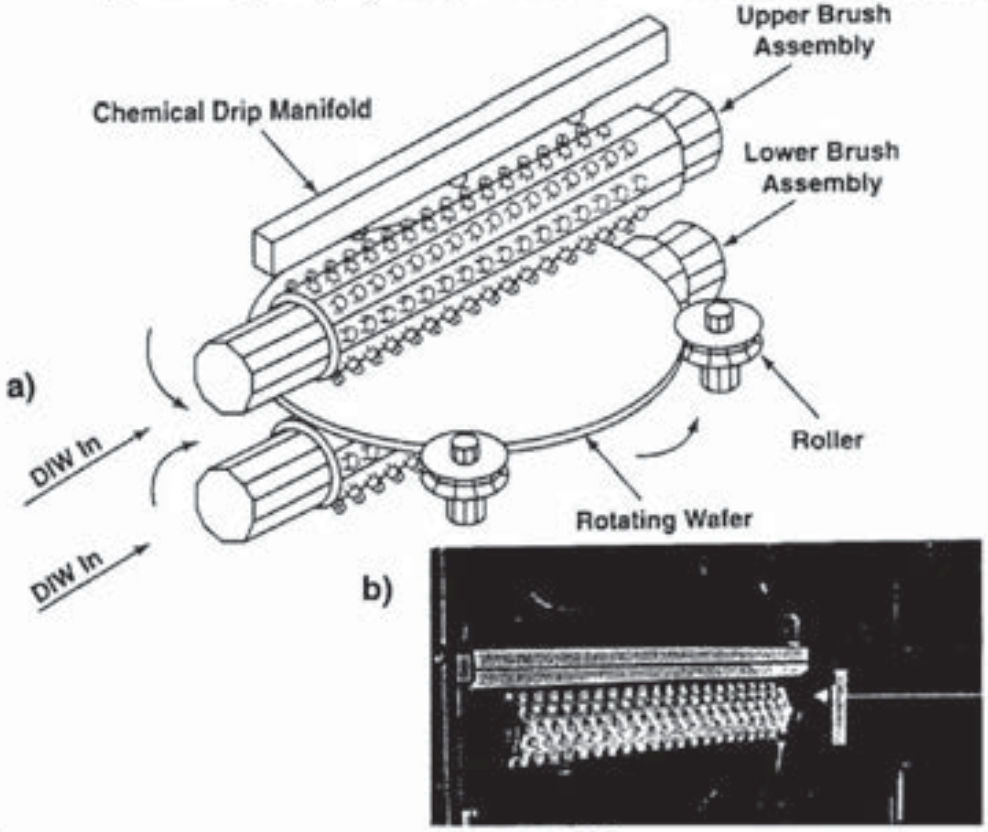


Fig. 8-80 a) Schematic of a double-sided brush scrubber.[65] Courtesy of Lam Research; b) Photograph of a PVA brush. Courtesy of Solid State Equipment.

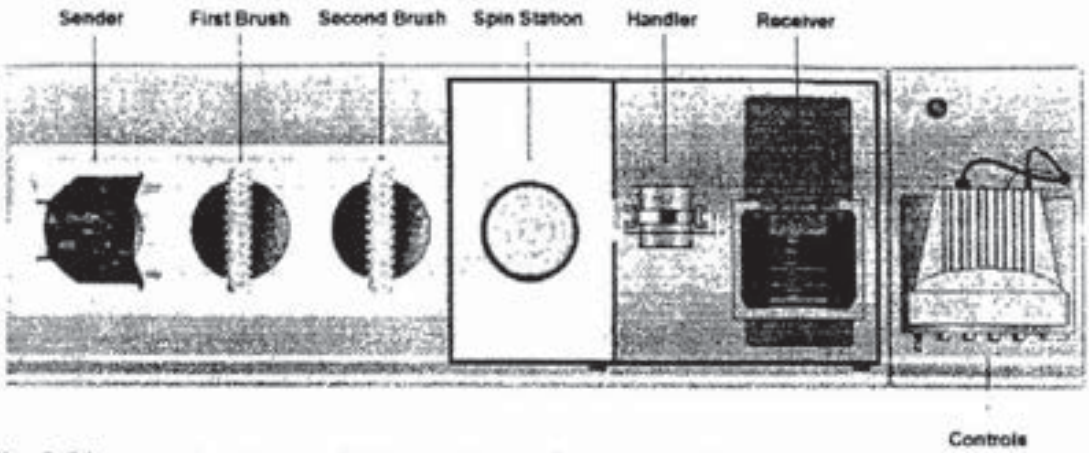


Fig. 8-81 Example of a post-CMP cleaning tool. Courtesy of Lam Research.

than 8. The post-CMP cleaning tool typically consists of a load station, a brush scrubbing module (or modules), a rinse-and-dry module, and an unload station (Fig. 8-81). Isopropyl alcohol and a heat lamp may be used to speed up drying.

8.7.2.1 Brush Scrubbing to Remove Particles in Post-Oxide-CMP Processes: When brush scrubbing wafers after an oxide-CMP process, a cylindrical polyvinyl-alcohol (PVA) brush (with or without nodules that protrude from its surface) is rotated across the surface of the wafer (Fig. 8-80), as the wafer is continuously flushed with the cleaning solution. The PVA brush-material is a soft, sponge-like material that is chemically resistant and highly compressible. This material allows the mechanical component of cleaning to occur without damaging the oxide surface. The brush is thus brought into direct contact with the wafer surface. Unlike rigid brushes such as nylon (commonly used in the 1970s), PVA does not cause scratches when brought in contact with the wafer surface.[63]

The PVA material is a sponge-like material that has an open structure consisting of interconnecting cells that allow the brush to be constantly flushed with DI water during wafer cleaning. These softer brushes clean more efficiently when they are compressed against the wafer. PVA brushes eventually wear out and need replacement. The average use rate is 20,000 wafers for post-oxide CMP cleaning, and 10,000 wafers for post-metal CMP cleaning.

It has been reported that higher brush speed can improve the particle removal efficiency without introducing scratches. For post-oxide CMP-cleaning (in which the CMP process uses silica-based slurries), the cleaning solutions use mixtures of DI-water and NH_4OH (and often H_2O_2, as well) to decrease silica-particle adhesion to the exposed oxide surface of the wafer.

One problem has been identified with respect to post-CMP cleaning, namely the difficulty of removing particles from non- or partially-planarized recesses. That is, as shown in the AFM image of Fig. 8-82,[64] a slurry particle resides inside a 0.7-μm contact hole with a 3:1 aspect-ratio after brush cleaning with dilute NH_4OH, and similar results were obtained with megasonic cleaning. It was observed that the ratio of the step-height to the horizontal-dimension of the feature is crucial for the cleaning capability of a specific cleaning technique.

8.7.2.2 Brush-Scrubbing to Remove Particles In Post-Metal-CMP Processes: In post-metal-CMP cleaning,, brush-scrubbing is also effective, but the process is modified from the one used for post-oxide-CMP brush-scrubbing. This is necessary because the post-metal-CMP process must: **1**) leave the surface of the metal without an oxide layer; **2**) leave the metal smooth; and **3**) not attack the metal. In addition, when cleaning wafers after a W-CMP-step, the wafers come from an acidic environment (2–4 pH) to a water-rinse and PVA-brush-clean. Some PVA materials cannot endure this chemistry well and must be frequently replaced. Other PVA materials (believed to be free of unreacted residues that hydrolyze in the acidic environment) appear to render them stable over an operating pH range of 2–12. These materials are reportedly compatible with post-W-CMP scrubbing.[66] It is reported that rinsing the wafers with dilute NH_4OH during post W-CMP scrubbing is also highly effective in removing alumina slurry-particles.[67] An additional benefit of using NH_4OH is that it prevents loading of the PVA brushes by alumina particles or tungsten oxides.

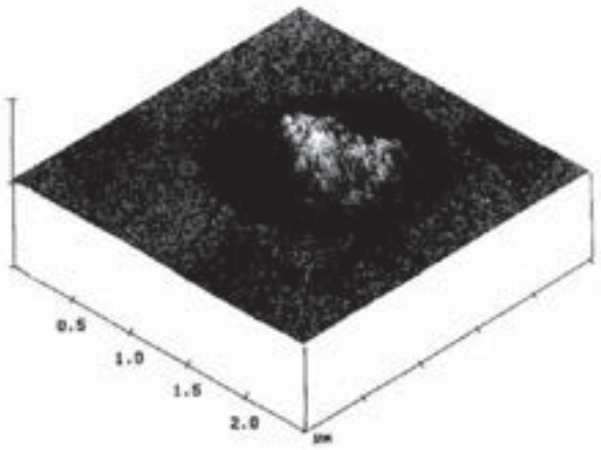


Fig. 8-82 AFM image of an agglomerated-silica slurry-particle in a contact hole.

Cleaning wafers after copper CMP presents serious challenges, including the following:

1. *The CMP-clean-step must not allow Cu-contamination of the fab cleanroom area*: That is, the Cu-CMP cleaning process must remove Cu from the front, back, and beveled edge of the wafer. However, the copper oxides formed and removed as part of the CMP process do not dissolve in DI water and may be easily transferred to the PVA brushes. If the brushes become "loaded" with these copper oxides, they may transfer them to subsequently processed wafers, thus producing severe Cu-contamination in the fab. Proprietary chemicals have been developed to dissolve the copper oxides and prevent such brush loading.[68] In the On-Trak systems brush-scrubbing tool, such a chemical is delivered to the wafer surface through the porous core of the brush itself (Fig. 8-83).

2. *The cleaning chemistry must not corrode the Cu*: If NH_4OH is used in the cleaning solution, great care must be exercised. This is because ammonia is known to complex with copper and cause surface roughness and corrosion products.[69]

3. *The Cu surface must be protected during the cleaning step*: Copper has a soft, porous oxide-layer through which chemicals can diffuse and cause further oxidation/corrosion. Protection of the copper-surface while cleaning the dielectric surface free of metals may be accomplished by using inhibitors.[70]

Note that more information on CMP of Cu is found in Chap. 16.

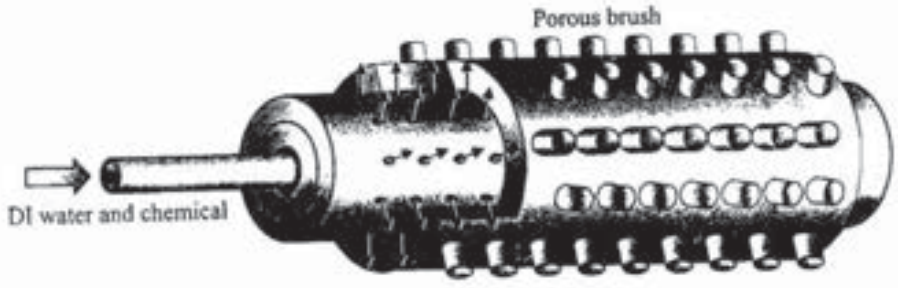


Fig. 8-83 *Through-the-brush* Chemical Delivery for post-CMP cleaning.

8.7.3 Vibrational Scrubbing to Remove Particles (Megasonic)

Brush scrubbing, while effective in removing particles, may cause scratching and brush shedding (which contributes additional particles). Vibrational scrubbing of wafers is an alternative non-contact and brushless technique for removing particles (Fig. 8-84a). Megasonic immersion cleaning systems for CMP using high-frequency sonic waves (0.7-1-MHz) are commercially available. The model of megasonic particle removal is shown in Fig. 8-84b. The wafers are immersed into an SC-1 solution contained in a megasonic tank. Sonic energy from acoustic transducers is directed parallel to the wafer surface, and this reduces the liquid boundary layer that surrounds a particle. The particle is thus subjected to larger fluid-acoustic-streaming velocities and energy from cavitation. These detach the particle from the surface.[71] The particles in the liquid must then be removed by passing the liquid through a continuous filtration system. This keeps them from remaining in the liquid where they could wander into the vicinity of the surface and be re-captured. Such filtration should also remove any particles generated by the megasonic system itself. (That is, particles shed by its own gaskets and transducer-bonding-material when the transducers are vibrating.)

The most common megasonic systems are batch-cleaning-systems. However, one recent report described a single-wafer megasonic-system that used a dilute basic-chemistry and cleaned each wager for 15-60 sec.[72] Cleaning with this system after an oxide-CMP-step yielded wafers with only 50 defects per wafer, well below the levels obtained when using brush scrubbing.

Megasonic cleaners may also offer cleaning improvement when topography such as alignment marks, scribe lines, or W-keyholes are encountered. That is, as discussed in the section dealing with *Brush Scrubbing for Oxide-CMP*, when particles are trapped within high-aspect-ratio recesses on the wafer surface, brush scrubbing is not effective in removing them. On the other hand, megasonic cleaning does a better job of dislodging particles in recessed features. Another advantage of megasonic cleaners involves cost. That is, cost is reduced since brush changes are not needed. The Applied Materials Mesa® combines brush scrubbing with a megasonic cleaner to remove particles in deep recesses. Other vendors, such as Verteq, also offer non-contact, non-scrubber CMP cleaners.[73]

8.7.4 Secondary (or Buff) Polish Procedure as a CMP Cleaning-Step

The polishing process itself can play a role in achieving low particulate levels at the conclusion of the CMP process. That is, by using a secondary buff-polish (i.e., polishing with a soft pad) and following this buff-polish with a high-pressure DI-water-rinse, fewer particles remain on the wafer after the subsequent

cleaning step. As a result, a second platen for buff-polish has become a standard feature for commercial polishers. It is also reported that if DI-water is used in the slurry (i.e., instead of a basic solution), particles counts are further reduced as the pressure during the buff-polish is increased.[74]

8.7.5 Post-CMP-Cleaning of Metallic Contaminants

Besides the removal of particles, post-CMP cleaning processes must also ensure that no metallic contaminants remain on the wafers when they are returned to the cleanroom area of the wafer fab. Since some commercial slurries are KOH-based, an effective clean must be employed to remove surface metals, especially alkaline metals. It has been reported that at least several-hundred angstroms of the top-oxide must also be removed to ensure that alkaline metal levels are adequately removed from the wafer-surface. Consequently, an "HF-deglaze step" is usually employed to reduce metallic contamination (and also any defects embedded within, or close to, the surface).

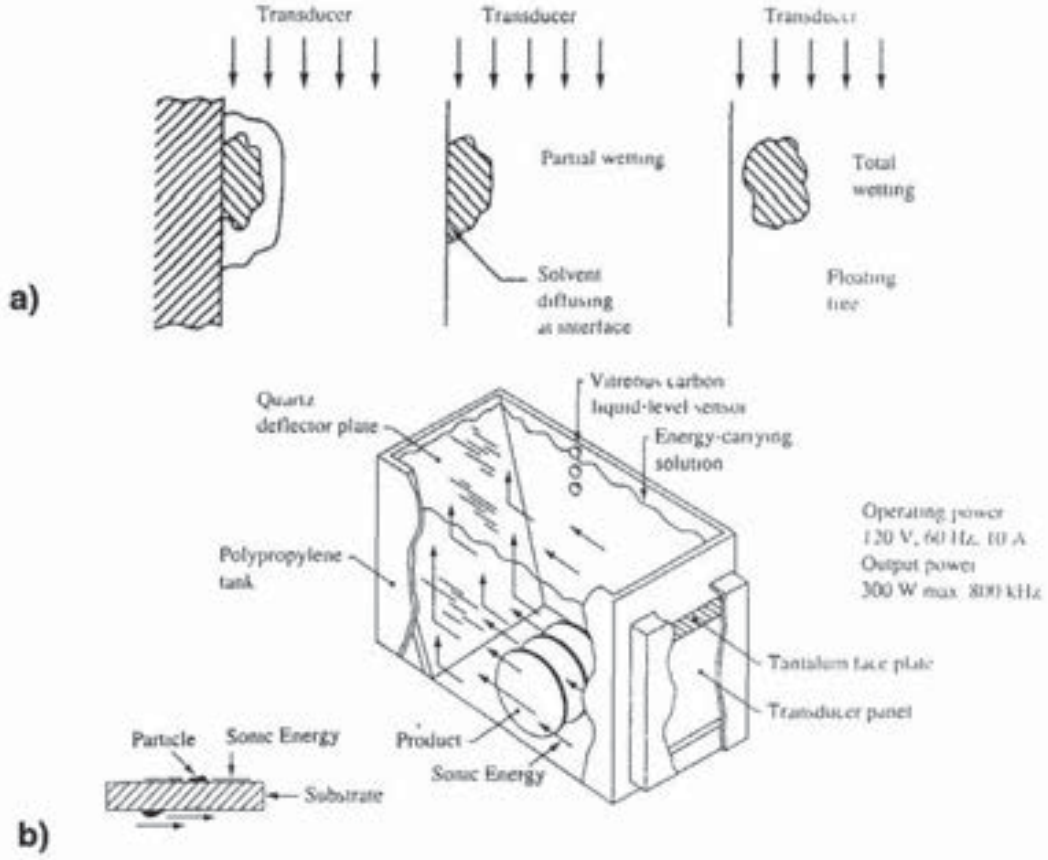


Fig. 8-84 a) Model for the mechanism of particle removal by megasonic scrubbing; b) Schematic of megasonic cleaning tank.

8.7.6 CMP Cleaning Equipment

On-Trak Systems (a division of Lam Research), supplies stand-alone double-sided scrubbing units for the post-CMP cleaning, as does Dainippon Screen. Other LAM post-CMP cleaning tools (e.g., the Synergy Integra® model, Fig. 8-85), are designed to be integrated with the polishing tools of other CMP vendors (including Applied Materials, SpeedFam, and Strasbaugh). Lam Research held about 70% of the post-CMP cleaning-tool market-share in 1999.

However, the Mirra® can also be integrated with an AMAT cleaning tool (the Mesa® Fig. 8-86).[75] It consists of four separate process modules: a megasonic bath; two-double-sided brush-scrub modules; and a spin-rinse-dry module. Each module cleans both sides of the wafer, and all handling is done by edge grip. Another fully integrated cleaner in a CMP tool is shown in Fig. 8-87.[76]

CMP processes are generally located adjacent to the cleanroom, connected through a bulkhead load, similar to etchers or CVD systems. In some cases, the CMP-tool-configuration consists of a polisher and stand-alone cleaning system with a robot that transfers wafers between them. A problem with this configuration being adjacent to the cleanroom is the potential for water and slurry spillage (and then leakage into the cleanroom) as transfers are made. A better solution is a *dry-in/dry-out CMP system*, in which the polisher and cleaning system are completely integrated. Wafers are planarized, cleaned, dried, measured, and returned back to the cassette located in the cleanroom.

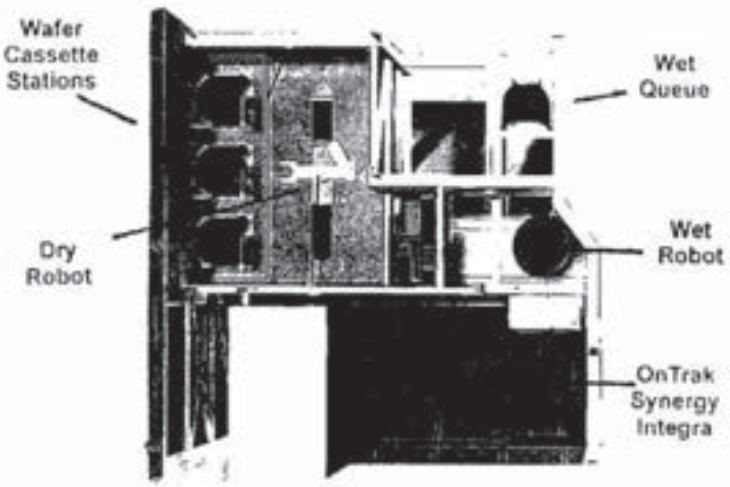


Fig. 8-85 Schematic of the Lam Research Synergy® post-CMP cleaning tool. Courtesy Lam Research.

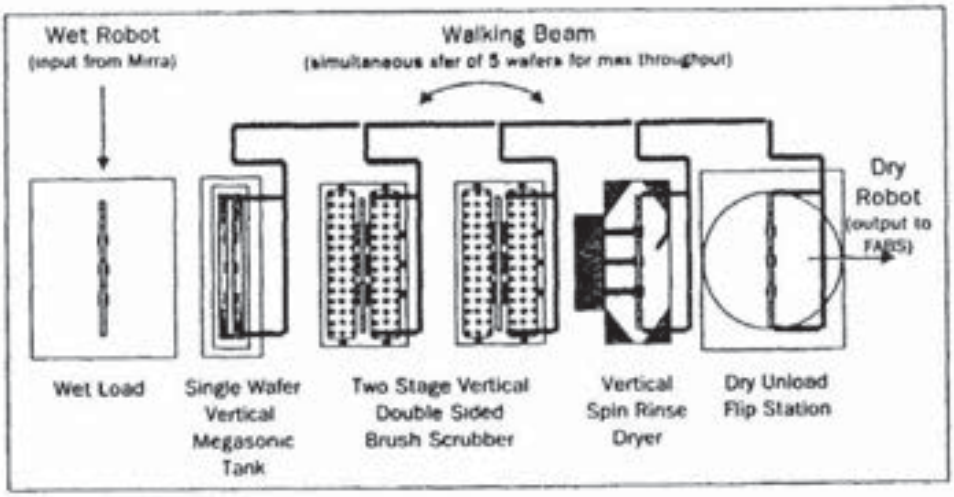


Fig. 8-86 Mirra Mesa® cleaning module layout. Courtesy of Applied Materials, Inc.

8.8 CMP METROLOGY

There are a number of metrology requirements in CMP. First, it must be determined when the CMP process is completed (*endpoint detection*, Sect. 8.5.3.4). Second, the film-thickness after CMP must be measured, including thickness-uniformity across a wafer. Third, the degree of film-planarity must be quantified after CMP. Finally, films subjected to CMP must be inspected for defects and damage. The latter two CMP metrology topics are discussed here.

Film thickness after CMP (and before CMP in in-line endpoint detection), is performed with a number of different techniques, including: spectrophotometry, ellipsometry, and beam-profile reflectometry. One problem associated with these techniques is that of *cycle-skipping* (or *order-skipping*). Since the signal obtained from the thickness-measurement is periodic, the same measured-data can fit more than one thickness-value. For oxide, the resulting errors are multiples of about 240-nm, enough to lead to incorrect conclusions about the post-CMP oxide thickness (which might cause properly processed wafers to be scrapped). Different strategies by each measurement technique are employed to avoid such problems. Ellipsometers use additional wavelengths, spectrophotometers rely on better signal-to-noise data-collection, and beam-profile-reflectometry is combined with spectrophotometry to avoid cycle-skipping errors.

One factor that impacts *thickness uniformity after CMP* is the uniformity of the wafer itself. It is observed that the *within-wafer-nonuniformity* (*WIWNU*) of the film removed by a CMP-process increases linearly with an increase in the

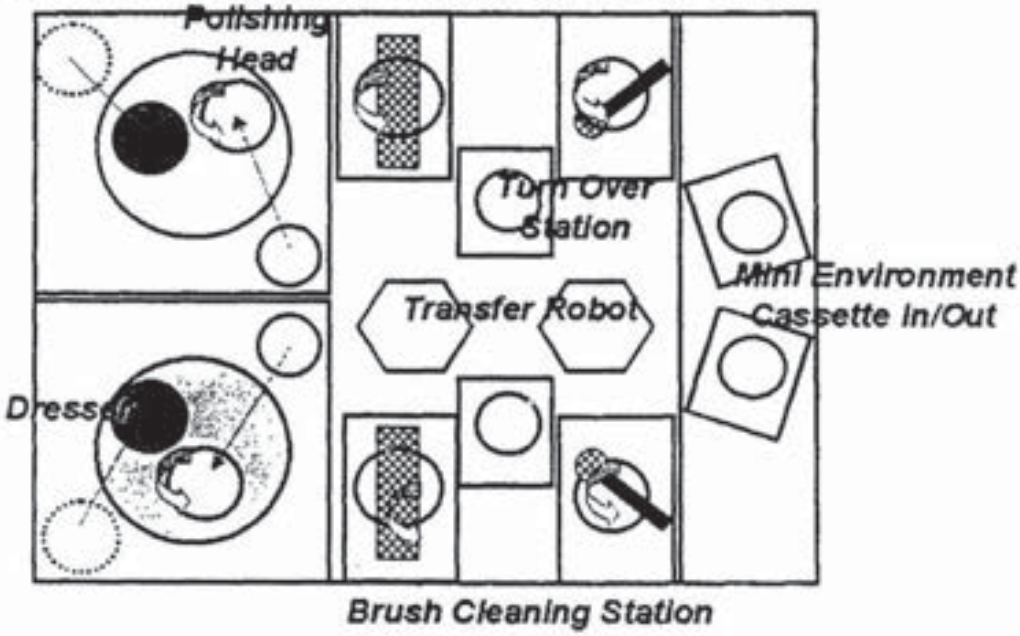


Fig. 8-87 Ebarra CMP cleaning module. Courtesy of Ebara.

total-thickness-variation, TTV, of the wafer.[77] In addition, wafers may not retain their initial flatness as they undergo the various steps of the process sequence. In particular, stress arising from thin-film-deposition processes can deform a wafer into a concave-shape (negative bow) or a convex-shape (positive bow). Although bow is reduced by the pressure applied to the wafer during CMP, some remains (as the degree of bow can vary from wafer-to-wafer). In addition some residual local-warpage persists after the wafer has been pressed against the wafer platen. Thus, the CMP-process may polish the edges faster than the center in a concave wafer. Wafers with lower pre-polish-bow reportedly had a smaller WIWNU than wafers with higher pre-polish-bow. As a result of the above issues, measurement of the wafer TTV as received from the wafer suppliers has been recommended.[77] Some CMP users have set an empirical TTV-cut-off-value of 5-μm for 200-mm-wafer pre-polish-sorting. Capacitive-gauging-tools are used to perform such TTV and bow measurements. More than 16,000 data points can be collected on a 200-mm-wafer in about one minute, and the results plotted to give a high-resolution map of the wafer geometry. Such gauging-tools are available from ADE Corp.

The *degree-of-planarity* achieved with a CMP-process is typically measured on a single-die or a few dice on a wafer with profilometry or atomic-force-microscopy. Such planarity can be location- and pattern-dependent. Within-die-planarity is influenced by the tendency of CMP to polish smaller, individual-features faster, and larger and densely-packed-features more slowly. Thus, CMP-thickness-measurements made with optical-thickness-tools commonly include

several features at each selected site (from the smallest that the tool can conveniently measure, to a large [field] area where the polishing-rate reaches its lowest limit).

8.8.1 Detection of Defects Associated with the CMP Process

Detection of defects on post-CMP wafers is another important aspect of CMP metrology.[78] This importance arises from the fact that CMP is probably the most yield-impacting of all IC process steps given its intrinsically "dirty" processing environment (which includes chemically corrosive slurries, nano-particulate abrasives, and polish by-products). The defects associated with the CMP process can take the form of: **1**) *damage*; **2**) *contamination*; and **3**) *particles*.

The acceptable density of defects is becoming lower for each device generation. In 0.18-μm CMOS-technology the SIA Roadmap calls for a defect-density of 0.14-defects/cm^2, and the minimum-detectable-particle-size is decreased to 0.06-μm from 0.08-μm for the 0.25-μm generation. In oxide-CMP, defects can be separated into two categories: **1**) *foreign material* (such as residual-slurry, surface-particles, embedded-particles, and residual-metal); and **2**) *voids in the surface material* (such as microscratches and dishing). *Microscratches* can occur when a particle is dragged across the wafer surface, causing a shallow void in the oxide layer. Figure 8-88 shows the defects arising from oxide-CMP processes.[79]

In metal-CMP, defect types include: metal-loss, puddles (which can cause shorts if they remain between metal lines), surface-voids, residual-slurry, surface-particles, metal-filled-micro-scratches (which can cause shorts), and plug-holes from which metal has been removed. Figure 8-89 depicts a variety of defects associated with the W-CMP-process.[79]

The detection signals caused by such defects are obscured to some extent by noise sources associated with the CMP process. The three prevalent noise sources that impact defect inspection at post-CMP processing are: **1**) layer-thickness-variation; **2**) grain-noise; and **3**) pattern-noise. *SiO_2-layer-thickness-variation* is caused by nonuniform-polishing across the wafer. *Grain-noise* arises from the surface-roughness caused by the polycrystalline structure of materials beneath the oxide being polished (e.g., tungsten, aluminum, and polycrystalline-Si). The noise occurs from randomly-distributed pyramidal-peaks at the surface of the poly-crystalline-films. These scatter the detection-signal in a non-consistent manner so it is not completely subtracted when two adjacent-die are compared. *Pattern-noise* arises from metal-features beneath the oxide (which distort the incoming signal by causing background-scattering). Defect-detection-tools

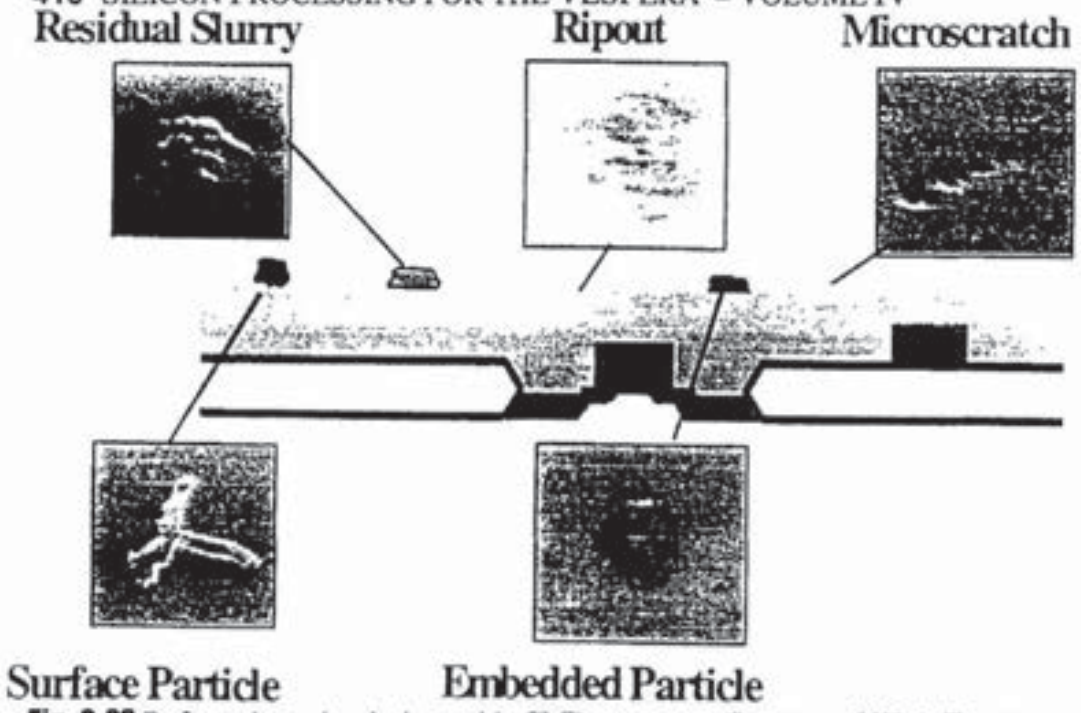


Fig. 8-88 Defects that arise during oxide CMP processes. Courtesy of KLA-Tencor.

employ image-processing algorithms to limit false alarms caused by such noise (i.e., that would indicate the presence of a defect where none actually exists).

Two illumination modes (*bright-field* and *dark-field*) are commonly used in the laser-scattering defect-counters employed for in-line automated-defect-inspection (such as the KLA-Tencor Surfscan AIT® model). Since reflected light is used in *bright-field inspection*, one of the main challenges for CMP-metrology is to reduce the interference between reflection at the oxide-surface and reflections that arise from underlying features. Some inspection tools use an ultra-broadband illumination-source (in place of a monochromatic-source) to suppress such interference effects. In *dark-field systems* only scattered light is collected, while the spectrally reflected light is rejected. Dark-field systems with a low-angle of illumination appear to work best for detecting slurry-particles on an oxide-layer over a grainy sub-layer (Fig. 8-90).[79] Dark-field systems with optical repetitive-pattern-filtering are reportedly more effective for detecting defects over dense repetitive arrays.[78]

Some 0.18-μm device-inspections for detecting CMP-defects may require use of SEM technology (such as the KLA-Tencor 2710). Because of their high resolution and large depth-of-focus, SEMs can locate ultra-small-defects hidden between dense geometries or in high-aspect-ratio areas. Small CuO particles ~0.1-μm from a dual-damascene process are easily detected using SEMs.

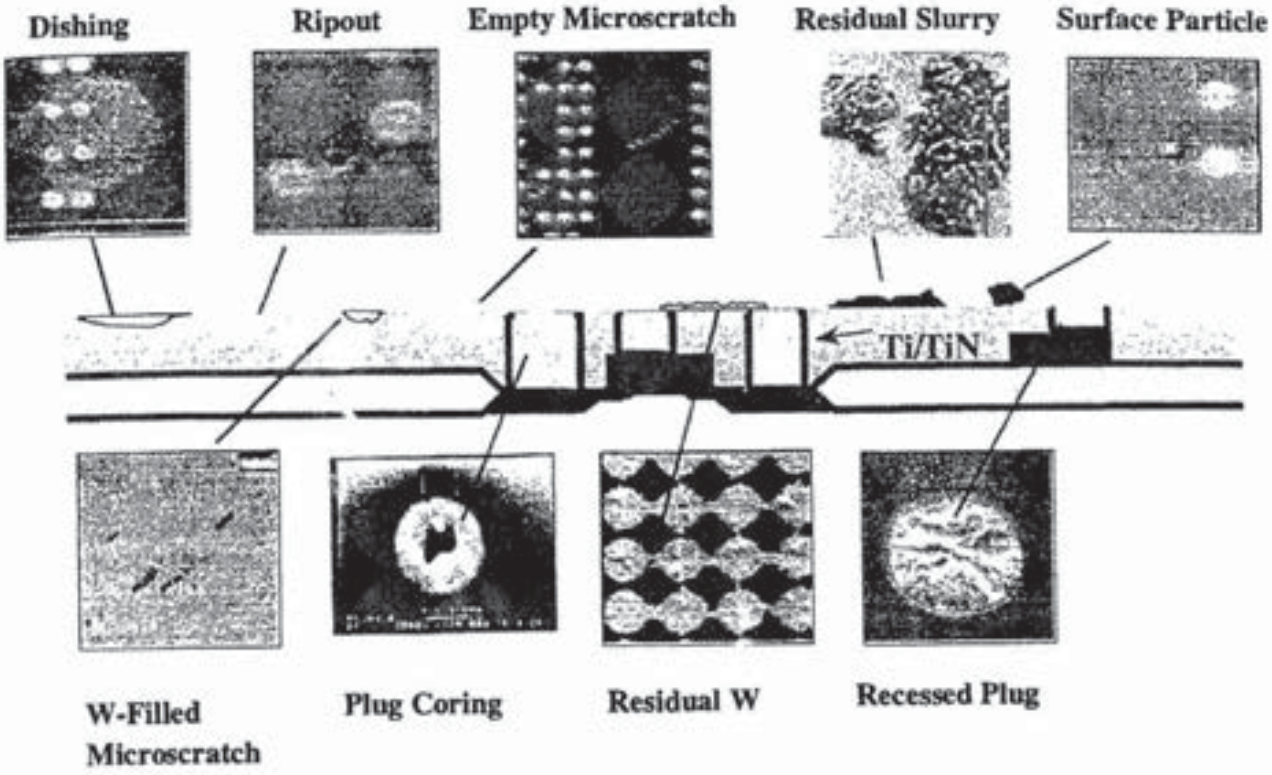


Fig. 8-89 Defects that arise during W-CMP processes. Courtesy of KLA-Tencor.

8.9 CMP POLISHER TOOL RELIABILITY

The CMP process creates a harsh set of working conditions for the polishers. Some have compared the operation of a CMP-polisher to "flying a helicopter in a sandstorm."[80] In the early days of CMP, production polishing-tools could have a mean-time-between-failure (MTBF) as small as a few days. Some of the problems of early CMP tools included the following:[39] insufficient vacuum for wafer handling; wafer sensor failure, misalignment of the load cup and sender and receiving cassettes, wafer release on polish pads (resulting in wafer breakage), equipment hangup (causing slurry to dry on the wafer-surface, and leakage of slurry and de-ionized water because of lack of quality control on valves and rotating unions. To make the systems more reliable, tools were designed to protect moving mechanisms from the abrasive and corrosive slurries (including sealing of parts and using air and water labyrinths).

Exposed materials were made with corrosion-resistant materials. Improved robotics for wafer loading, unloading, and handling were implemented. New system components were subjected to accelerated-testing to simulate one-year of

production in three months. As a result, MTBF rates in CMP tools have been continuously improving. The goal of CMP tool suppliers is to build polishers with a MTBF of at least 400 hours (about 16 days).

8.10 CMP SYSTEMS AND PROCESS INTEGRATION

The CMP process requires a host of tools and consumables (Fig. 8-91). When CMP first came to the IC industry, each component of the CMP process was supplied by a separate company. The burden of integrating the various components fell to the user. As the market for CMP tools and consumables grew, equipment and materials suppliers realized the need for a better CMP-infrastructure. They understood that to gain market share, CMP users should be offered an off-the-shelf, "one-stop-shopping" package, perhaps with guaranteed process results. One way to accomplish this was to form strategic partnerships with other (non-competitive) suppliers. Most polisher suppliers included slurry distribution systems, cleaning tools, pads, slurries, endpoint detection, and film thickness measurement tools from other vendors as available options compatible with their polishing tools. Some companies (including Lam Research, Applied

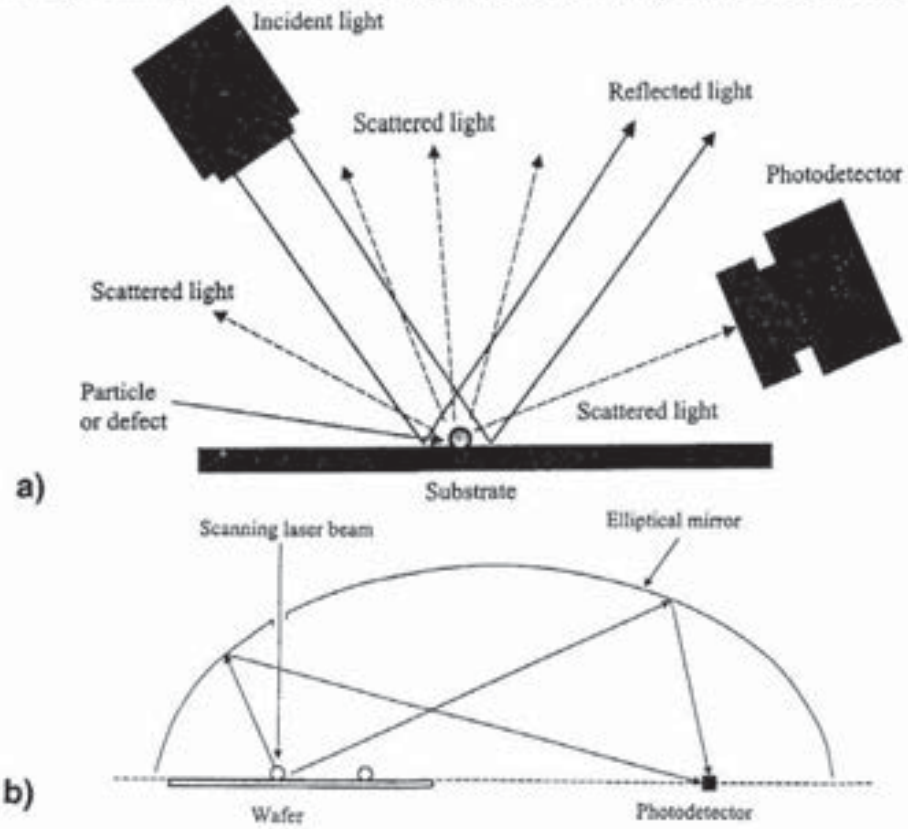


Fig. 8-90 a) Particle detection by Darkfield Light-Scattering; b) Light-scattering particle detector with an elliptical mirror.

Materials, SpeedFam-IPEC, Ebara, Strasbaugh, and Peter Wolters) designed CMP systems as cluster tools. Such systems integrate polishing, endpoint detection, cleaning (and perhaps inspection) into a single system. The Lam Research Teres® and the Applied Materials Mirra Mesa® are examples of such cluster CMP tools.

These integration trends have been accompanied by business consolidations as well. Speedfam and IPEC (two large polishing tool suppliers) merged in 1999. Lam Research acquired On-Track Systems (a maker of polishing tools and post-CMP-cleaning-systems). Applied Materials announced it would acquire Obsidian, Inc., a company that had developed a fixed-table CMP tool which uses a slurry-less pad technology. As CMP technology matures it is likely that such consolidations will continue.[82]

8.11 CMP OF VARIOUS MATERIALS

Here the issues of oxide-CMP and W-CMP are described in detail. The subject of Cu-CMP is briefly introduced in Sect. 8.11.3 here, but more details on Cu CMP are provided in Chap. 16 (that deals with the process technology of Cu in ULSI).

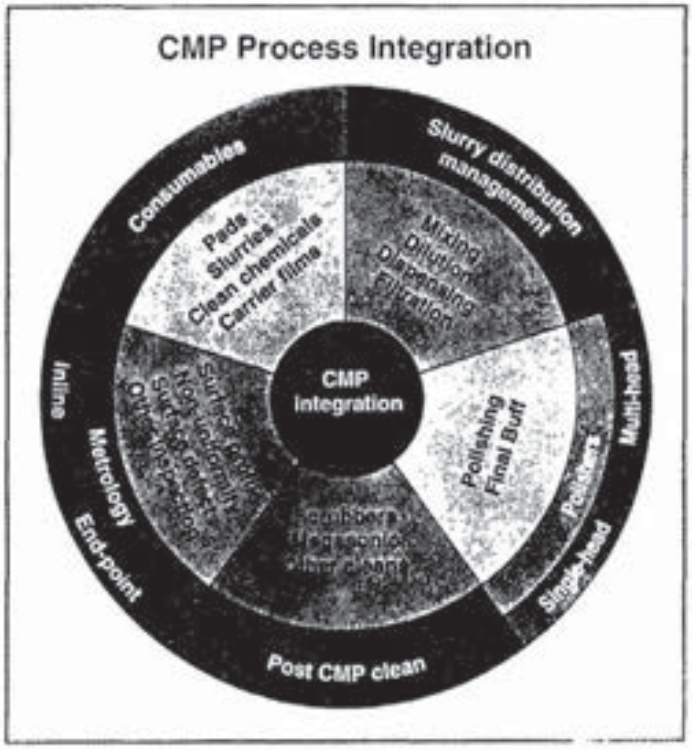


Fig. 8-91 Many components of CMP must be successfully integrated to develop a production-worthy CMP process. Reprinted with permission of Semiconductor International.

Note that the CMP issues associated with shallow-trench-isolation (STI) are mentioned briefly in this chapter in Sect. 8.12 (which deals with the WIDNU phenomenon of *dishing*). However, additional details concerning CMP-issues in STI are found in Chap. 9 (which is dedicated to the topic of STI).

8.11.1 CMP of Oxide Interlevel and Intermetal Dielectrics (ILDs and IMDs)

A blanket-dielectric-layer (typically an oxide-like film) is deposited (generally by CVD) atop polysilicon after gate-patterning. This layer is called an *interlevel dielectric* (ILD). Likewise, a blanket dielectric-layer (again, usually of silicon oxide) is CVD-deposited onto patterned metal-features, and such layers are called *intermetal dielectrics* (IMD). Topography is produced when oxide films are deposited by CVD onto etched polysilicon and metal patterns (see Fig. 8-1a). Many techniques have been attempted to planarize this topography (including: reflow for ILDs; and resist-etchback and spin-on-glass for IMDs). CMP has replaced these planarization techniques because they can only provide local-planarization (see Fig. 8-1d), whereas CMP can provide global-planarization (see Fig. 8-1e). However, unlike CMP for STI-applications, there is typically not a polish-stop-layer, and repeatable control of the polish-rate and uniformity is required. As noted previously, process variations (such as pad-age, slurry batch-changes, machine-calibration-drift, pad-conditioner-age, and chamber-to-chamber dielectric-film-property-variations) can all result in significant removal rate and process non-uniformity.

The industry standard pad for ILD and IMD CMP has been the IC-1000/Suba-IV stacked pad. This pad consists of a hard-cast polyurethane-pad (IC-1000) on top of a soft, compressive felt-pad (Suba-IV). The hard-top-pad is needed for planarization, and the soft-pad underneath improves the WIWNU (compared to polishing without the soft underpad). The drawback for including the soft underpad is a lower degree of planarization. Total stack compressibility can be modified by varying the thickness of the two pads.

Another issue of ILD-CMP is that doped silicon-oxides, such as PSG, and BPSG are softer glasses that undoped silicon-oxides (USG). As a result, they exhibit faster removal rates than USG. Furthermore, doped ILD-films exhibit higher potassium and calcium levels on the wafer-surface after scrubbing and cleaning than do USG films. This occurs because the phosphorus in the PSG and BPSG films acts as a gettering species for metal-impurities in the oxide-film. An HF-dip in the post-CMP-cleaning -equence can reduce the level of such metallic-contamination on the PSG surface (as well as slurry-residue). However, soft films, such as PSG tend to have high post-CMP defect-counts, and the HF etches

microscratches isotropically, and enlarges their size. Thus, the post-CMP-cleaning-processes of PSG and BPSG ILDs need to be optimized to simultaneously reduce metallic-contamination and defectivity (e.g., by performing an oxide buff, which is very effective in reducing microscratching).[82]

8.11.2 CMP of Tungsten

Tungsten (W) CMP has been developed to replace the W-etchback step of the W-plug formation sequence (see Chap. 6, Vol. 1, 2nd Ed.). The W-CMP process, however, requires a surface that is planar prior to via-formation and metal-deposition (a condition not required in processes that form W-plugs by W-etchback). W-CMP has been described as a balance between the electrochemical oxidation of W to form a blanket passivation layer WO_3 on the entire W surface, and the mechanical removal of that reaction-product by abrasive slurry particles. Tungsten, being one of the refractory metals, tends to form strong oxygen bonds under oxidizing conditions. This oxide is so dense that the metal is protected from further chemical attack and the metal is "passivated."

Polishing of tungsten films raises three key issues. The first involves selection

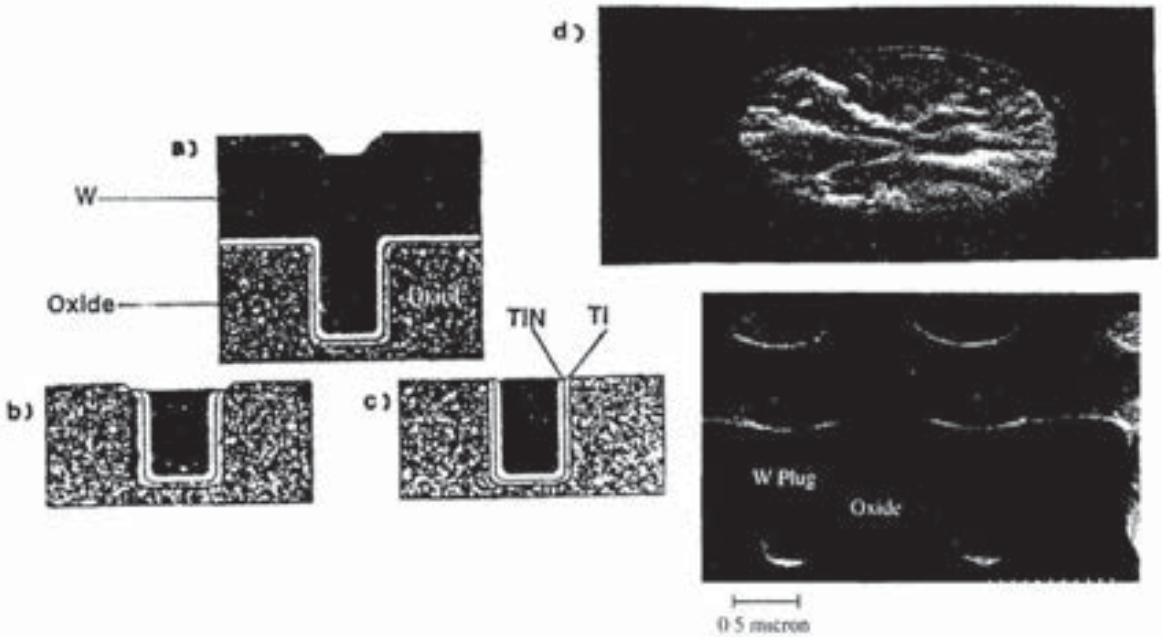


Fig. 8-92 W-plug formation using CMP: a) W deposition over TN/Ti barrier layers; b) CMP of W and TiN/Ti leaves recessed plugs; c) Oxide buff eliminates recess.[84] Reprinted with permission of Solid State Technology. Published by PennWell.; d) SEMs of W-plugs after W and TiN/Ti CMP steps and oxide buff step.[85] Reprinted with permission of Semiconductor International.

of an appropriate slurry. Tungsten polishing at IBM was originally done with a slurry-mixture of potassium ferricyanide and potassium phosphate (having a pH adjusted to about 5.0 to 6.5), together with colloidal-silica or dispersed-alumina, Al_2O_3.[83] (Alumina-based slurries have come to be the more commonly used.) However, potassium ferricyanide is highly toxic, and its disposal causes severe environmental impact. As a result, this chemical has been replaced, with the most widely used substitutes being ferric nitrate $\{Fe(NO_3)_3\}$, potassium iodate (KIO_3), and hydrogen peroxide (H_2O_2).[84] An example of commercially available slurries for W-CMP are Cabot's SEMI-SPERSE® WA-355 and SEMI-SPERSE® FE-10.

The second key issue of W-CMP arises because blanket-W-films are deposited over a liner-film of TiN/Ti. When removing the W-film, this liner-film must also be polished away. A high removal-rate of the TiN/Ti layer is essential to prevent erosion of the oxide underlying the metal-layers, and to also prevent plug-recessing (Fig. 8-92a). However, polish-rates of TiN/Ti are typically three to eight times slower than that of W (which is removed at rates exceeding 350-nm/min). It is also necessary to have a high selectivity to the underlying oxide to obtain local planarity of the plug-surface and to prevent excessive loss of the surrounding oxide by dishing. To remove the liner-film, *over-polishing* is typically used (i.e., so named because it causes some recessing of the W-plug as shown in Fig. 8-92b). During this step the field-oxide is also damaged to some degree.

To solve the problem of having to polish two-different metals, a multi-step polishing sequence is often used in W-CMP processes (typically a three-step CMP process is needed).[84,85] For example, in the first step, the W-layer is removed using one platen of a polishing tool. In the second step, the wafer is transported to a second platen where the liner-film (e.g., TiN/Ti) is polished away. Finally, a third oxide-buff-step (on yet a third platen) removes the oxide-damage by polishing away a small amount of the now-exposed underlying oxide (using a silica-based slurry that minimizes surface-roughness and facilitates the removal of alumina slurries). This also restores the level of the W-plug to that of the oxide-surface by eliminating the recess (Fig. 8-92c). Such CMP-W-plug-processes reportedly produce higher yields than processes which employ a W-etchback-process.[86,87] The higher yields with W-CMP are attributed to improved defectivity and superior final plug-morphology.

The third problem associated with CMP of tungsten involves the voids or seams that can occur when the blanket-W-film fills the contact or via holes. If such voids exist, they will be exposed during the CMP process. These voids become regions in which slurry can be trapped. Slurry chemistry can also attack a seam along a W-plug center and enlarge the seam to form a void or widen/expose

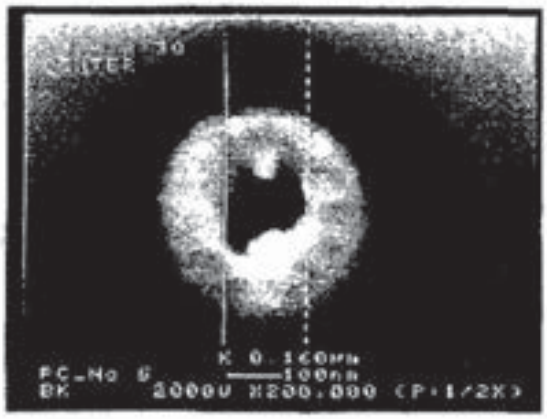


Fig. 8-93 Plug-coring.

an existing void, a phenomenon termed *plug-coring* (Fig. 8-93). Prevention measures include developing W-deposition-processes in which void-formation does not occur, and the use of less aggressive slurry chemistries (which will have less tendency to attack weak seams).

8.11.3 CMP of Copper

The CMP of copper (Cu) interconnect films involves a whole set of unique challenges,[88,89] and some of these issues are described here

The removal-rate of Cu by CMP is about 350-nm/min with a selectivity of 150:1 to the underlying oxide. However, Ta is commonly used as the liner-layer for Cu, and removing it by CMP represents some difficulties. Cu is relatively soft and oxidizes easily with standard chemistry (such as hydrogen peroxide or hydrogen persulfate). On the other hand, Ta is very hard and is also much more chemically inert. This typically leads to large differences in selectivity when simultaneously polishing Cu, Ta, and the underlying oxide (i.e., the Ta removal-rate is only about 30-nm/min). This leads to significant *dishing* of wide Cu and oxide features (the term *dishing* is defined in Sect. 8.12). Damascene CMP applications require low selectivity between the primary conductor-material and barrier-films (1:1) and a high selectivity between the primary conductor and insulator (> 100:1) to essentially stop the process. Since Ta is the harder and more inert metal, it is thought to be removed primarily by abrasion. An aggressive oxidizer could be used, but this would create the possibility of polishing through the Cu and removing device features. Work has been proceeding to find an oxidizer that selectively attacks the Ta without removing Cu.[90]

Another related problem in Cu/Ta-CMP is the electrochemical-interaction between Cu and barrier-materials during CMP.[91] Two dissimilar metals in

simultaneous electrical-contact and contact with an electrolyte-solution constitute a galvanic-cell. This chemistry frequently leads to anodic-corrosion. Such a situation arises near the end of a polishing cycle of Cu and Ta (when the barrier-metal becomes exposed in the field-regions, and Cu is exposed in the trenches). Cu in the trench can be removed (sometimes entirely), even if the trench is initially well-filled with Cu. Obviously, this effect must be suppressed to enable CMP of Cu/Ta films to be successfully performed.

Use of multi-step polishing-processes has also been pursued for CMP of Cu. As in W-plug-CMP, a three-step damascene-Cu-CMP process uses three platens, each with its own slurry mixture. Copper is removed with the first platen, Ta layers with the second, and an oxide-buffing step is carried out on the third. It must be ensured that all of the Cu in the field-regions is removed during the first step, since the removal-rate of Cu in the second is deliberately kept very small. In another approach used by Speedfam, multiple slurries are simultaneously delivered through-the-pad to each of four autonomous polishing modules. In this way, a multi-step Cu CMP process can be carried out without transferring wafers from one platen to another (which increases throughput and yield).

For multi-step Cu-CMP processes one of 3 or 4 types of slurry might be best for the second polishing step. The selection of a particular slurry can depend on the liner-layer being used and the type of ILD material beneath the metal film. A large number of companies are developing slurries for Cu-CMP, hoping to capture a significant market share of this important emerging process.

8.11.4 CMP of Aluminum

CMP of aluminum is being examined for use in damascene-Al-interconnects. One of the challenges facing Al-CMP is that the metal is soft and susceptible to scratching, but its oxide Al_2O_3 is hard and adheres tightly to the metal. This makes Al_2O_3 difficult to remove. If a strong oxidizer is used, the entire Al-film (not only its elevated regions) tends to disappear. It has been shown that if a correct set of process conditions is established (e.g., an appropriate combination of pad, slurry type, and applied pressure), Al films can be planarized with CMP. While removal-rates of about 200-nm/min have been demonstrated, resolution of the issues concerning the Al-CMP process had not yet occurred by 1997.[92]

8.11.5 CMP of Polysilicon

Planarization of polysilicon layers by CMP is carried out in some IC fabrication applications. These include the following:

1. Removal of the polysilicon from the wafer surface when a blanket-polysilicon-film is deposited to fill the trenches of trench-capacitor based

DRAMs. (The issues of poly center-seam attack observed in some RIE etchback-processes used to remove the poly do not occur if a poly-CMP process is used. In addition, the polysilicon slurries are highly selective to oxide or nitride hard-mask materials and thus, are self-stopping);

2. Topography that exists when a polysilicon-film is deposited over an STI structure may need to be planarized. As shown in Fig. 8-94, the topography near the edge of an STI structure is often not perfectly planar. Since the polysilicon deposition process is inherently conformal, the non-planar surface of the STI-structure is replicated in the overlying poly film. In some applications it is advantageous to remove such undulations.

Polishing of polysilicon-films is done with a CMP-process similar to that used to planarize oxide.

8.11.6 CMP of Photoresist

There are some applications where the ability to CMP photoresist (PR) is beneficial. One example involves the fabrication of trench-based DRAM capacitors in embedded-DRAM technology. In the fabrication of the deep trenches used in such DRAM-capacitors, a sacrificial PR-layer is spun onto the wafer to control the height of the surface on the wafer and to facilitate alignment when building other device structures in the vicinity of the trench array. Both within-wafer and within-die non-uniformities of such sacrificial-PR-layers are important, as schematically shown in Fig. 8-95. These PR layers were found to

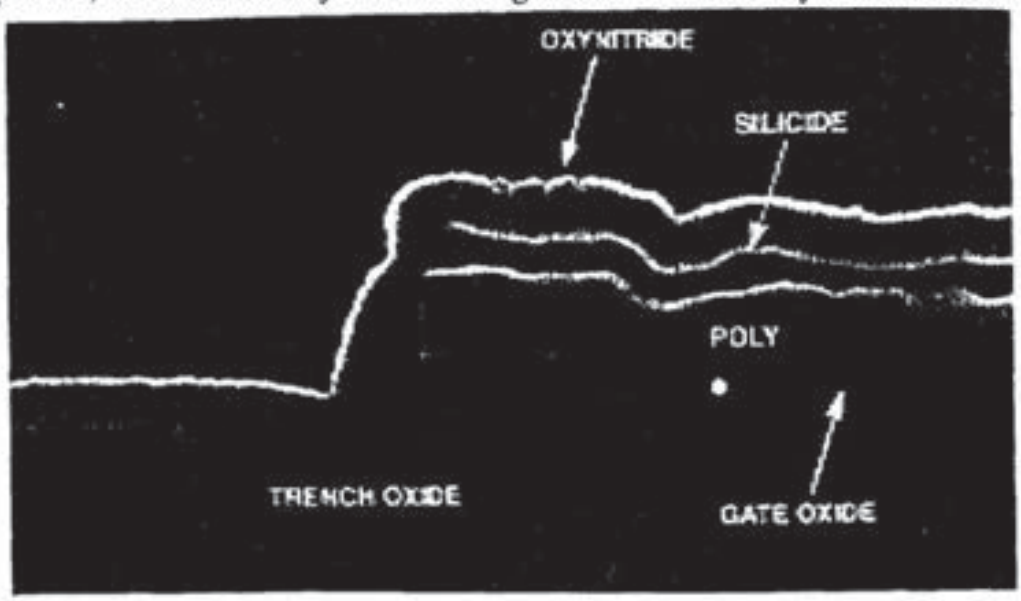


Fig. 8-94 CMP of polysilicon.

be successfully polished with a low-cost CMP process that utilized a slurry-free pH compensated solution of DI water (i.e., no slurry particles).[93]

8.11.7 CMP of Low-*k* Dielectrics

Other materials besides CVD-SiO_2 are being investigated for use as intermetal dielectric films (especially those with a lower dielectric constant - low-*k* dielectrics). These materials are discussed in detail in Chap. 14, but here the issues of low-*k* dielectric CMP are considered. That is, to be used in damascene interconnect structures, such low-*k* materials must be compatible with CMP.

Many of the CMP issues involving low-*k* dielectrics are common to both the subtractive-Al and damascene interconnect schemes. Most of the low-*k* stack options intend on keeping the low-*k* material encapsulated, never exposing it to the CMP process and post-CMP clean. In these cases, the investments in processes, materials, and equipment for silicon-oxide-based CMP systems will carry over and maintain their value. Film-adhesion is fundamental to all on-chip materials, particularly during CMP, with high friction and shear forces at work. Most proposals include capping the low-*k* dielectric with a traditional layer of silicon-oxide (or in some cases, with silicon nitride). The low-*k* materials must adhere well to both the cap-layer on top and to the underlying layer below (which is also likely to be silicon-oxide or silicon-nitride). In addition to adhesion, the compressibility of low-*k* films will make it possible for CMP processes to cause some local flexing within the interconnect stack. Such flexing could lead to cracking of the more rigid cap layers and produce localized delamination. In addition to unplanned exposure of the low-*k* film to slurry and post-CMP cleaning chemicals, the cap fragments could lead to scratching. Consequently,

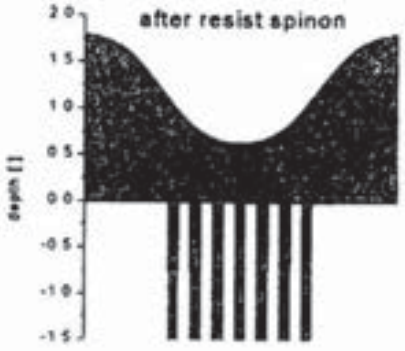


Fig. 8-95 CMP of photoresist.

polishing of cap layers on low-*k* is one of the earliest tests performed before a new low-*k* material can be considered integration-worthy. Yet another low-*k* CMP problem involves the low-thermal-conductivity of such films compared to silicon-oxide. Frictional heating in the CMP process may give rise to hot spots, possibly resulting in localized-stress and cap-delamination. High-linear-speed equipment being developed for greater planarization efficiencies would exacerbate this problem. In the extreme, local temperature excursions may further cure the low-*k* film, resulting in non-uniform dielectric properties.

Up to the end of 2000, there was little published information dealing with the issues of carrying out CMP of low-*k* materials in a volume-manufacturing environment. Part of the difficulty lies in the fact that process and consumables requirements may vary widely, depending on the specific low-*k* material chosen. One published report on low-*k* dielectric CMP involves the CMP of fluorinated oxide (F-SiO_2).[94] It was found that both mechanical-abrasion and chemical-erosion are intensified by fluorine-doping (relative to undoped TEOS oxides). This happens because of the reduced hardness and chemical reactivity of fluorinated oxides. One way to avoid these problems is to cap the F-SiO_2-layers with a CVD-SiO_2-layer. This cap-layer must be thick enough so that the dielectric-stack is planarized by CMP. In such cases the F-SiO_2-layer is not reached. Dielectric-erosion due to CMP-dishing must also be less than the thickness of the cap-layer. Furthermore, the cap-layer increases the overall dielectric-constant of the IMD-stack. In addition, depositing the stacked-film is a multi-step-deposition process. Another approach is to pursue the development of more chemically-stable low-*k*-materials, and such efforts are continuing.

A second study of direct polishing of low-*k*-dielectrics using a MnO_2-slurry reported that the deep scratches formed with CMP using a conventional silica-based slurry were reduced from a 20-nm depth to a 2-nm depth. Residual MnO_2 slurry particles can also be removed using H_2O_2 with a pH of 6, which does not attack the fluoro-carbon low-*k* material being polished (a poly-AE layer, FLARE®).[95]

8.12 MISCELLANEOUS ISSUES CMP

8.12.1 Dishing and Erosion

Dishing, another CMP issue, can represent a significant problem when a film of one type of material is being polished down until it remains only in recessed (or trench regions - Fig. 8-96). In such cases the layer under the film being polished

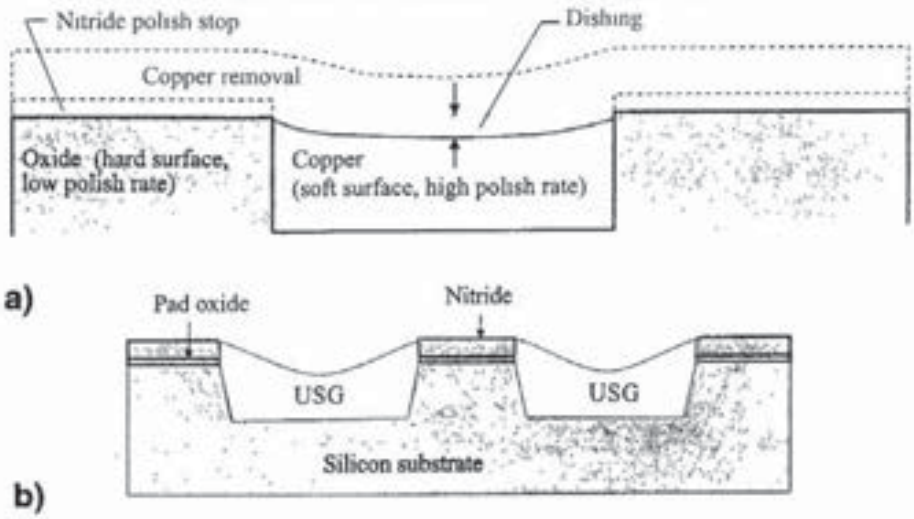


Fig. 8-96 Dishing-effect, shown here for: a) CMP of C in oxide-trenches; b) CMP of oxide in Si-trenches during formation of STI.

becomes exposed to the CMP process during the over-polish time (used to ensure that the film is completely removed over all regions of the wafer by the CMP process). Ideally, such underlayers should act as perfect polish-stop-layers. However, during the over-polish time, the film being planarized can continue to be polished away (especially in wide areas), causing it to be significantly thinned in such areas. One case where this occurs is in the CMP of a CVD-oxide-layer, deposited to fill trenches in the course of fabricating shallow-trench-isolation (STI) structures. A silicon-nitride-layer is typically selected to act as the polish-stop-layer. (See Chap. 9 for a discussion of STI technology.) During the over-polish time, the oxide in the trenches continues to be removed. The degree of thinning of the oxide in such areas (*dishing*) is known to be highly dependent on feature size, with the dishing-effect becoming larger as the width of the field-oxide is increased (Fig. 8-97). For example, in some processes dishing can cause the field-oxide thickness to decrease by more than 50-nm when the field-oxide width exceeds 100-μm.[96]

Dishing occurs because the polishing-pad is not perfectly rigid. If the pad were to exhibit perfect rigidity, CMP of the exposed oxide-regions would produce a perfectly planar surface (because the pressure on the exposed recessed regions would be zero). Since actual pads exhibit some flexibility, above-zero polishing-pressures are produced in recessed areas, especially in the middle of wide exposed-regions. Thus, oxide is removed from these regions, and a "dished" region exists after planarization. Dishing must be minimized to maintain planarity for subsequent wafer processing. It has remained a serious impediment to developing damascene processes (as is *erosion*, which is covered later).

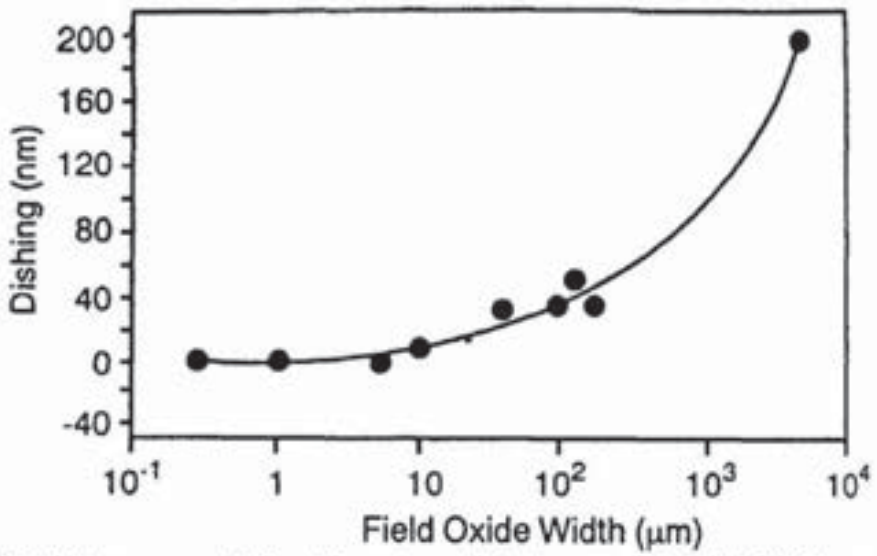


Fig. 8-97 Dishing versus field-oxide-region width for a specific oxide CMP process.

A number of approaches have been proposed to decrease dishing arising from the CMP-process used to fabricate STI, including: **1**) a combination CMP/RIE process; **2**) dummy structures added in the field-oxide regions of STI structures; **3**) adding an extra nitride-layer on the trench-dielectric after it is deposited;[103,104] **4**) adding an extra nitride-layer onto the dielectric after it has been deposited *and* using a reverse block-mask;[105] and **5**) implementing a direct-polish CMP-process using either a linear-polishing-tool, or a rotating-polishing-tool with a harder polishing-pad and an efficient endpoint-detection-system. Techniques #3 and #4 are described in Chap. 9 (STI), and the others here.

In the *CMP/RIE-approach*, the oxide is polished by CMP to a point where the nitride is not yet exposed and the surface is thus still completely planar (no dishing has yet occurred). Then RIE is used to etch the oxide-layer, with an endpoint-algorithm used to stop the etching just as the nitride is exposed to plasma.[97] The drawback of this method is that it increases the complexity of the process used to planarize the trench-oxide.

In the *dummy-active-region-approach* some areas of the trench are not etched. Instead, as shown in Fig. 8-98, regions of Si remain in the center of the trench. This essentially reduces the trench-width so that dishing effects are decreased to an acceptable degree. The CMP-removal-rate over regions that are densely packed with active-areas is also slower than over regions containing small isolated areas. Thus, to expose the nitride CMP-stop-layer over such densely packed regions, the nitride over small isolated active areas gets stripped, causing damage to Si in areas beneath the stripped nitride. Thus, dummy-active-regions

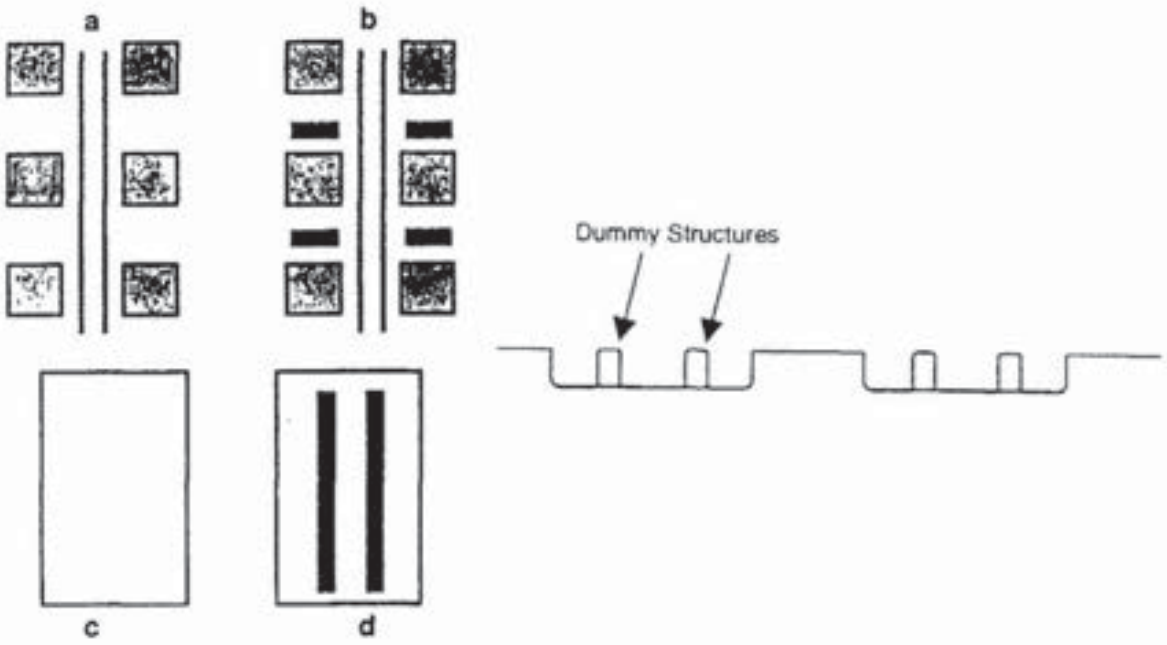


Fig. 8-98 Dummy (b) and dishing support features of small (a) and large wafer (c) features can prevent dishing.[89] e) Cross-section of dummy features in a trench.

must also be added to areas around isolated active regions. Finally, since dishing also occurs when wide areas of metal are polished (e.g., bond-pads, power-busses, and wide interconnect-lines), such dummy structures can also be added to these features. The technique, however, poses electrical, design, and space constraints, with the main issue being that several times as many structures must be added within a die over previous designs. This complicates the design and mask-making process, and significantly raises defect levels. Thus, some see this approach as merely a stop-gap measure.

The implementation of a *direct-polish CMP-technique* has also been pursued to avoid the drawbacks of the above methods. When a direct-polishing-process involves the use of a *linear-polishing-tool*, higher pad velocities (300–500-ft/min versus 34-ft/min for rotary-pad speeds) and lower polishing-pressures (3–5-psi versus 7-psi for rotary polishers) are used to suppress dishing. A highly-selective slurry (with cerium-oxide particles dispersed in an aqueous suspension of material that inhibits silicon-nitride removal) is also employed. This set of process conditions has been reported to significantly reduce dishing-effects (maximum oxide-removal across a die was 20-nm) and decrease the nitride-removal over isolated active-regions.[98] Decreased dishing using a direct-polish approach has also been described using *rotary-polishing-tools.* To achieve decreased dishing the correct set of process conditions and CMP consumables must be selected. That is, dishing effects are reduced by using a harder polishing

pad, a relatively non-selective slurry, an efficient endpoint-detection method, and an optimized CMP process.[99] As also discussed in Sect. 8.6.8, the use of slurry-free pads holds the promise of also reducing dishing.

The dishing problem also occurs during other CMP processes as well. As noted in earlier, dishing can occur when polishing ILD and IMD layers. Furthermore, as will be discussed in Chap. 16 (which deals with Cu interconnect processing issues), dishing and erosion of the ILD layers can occur when the Cu (and Cu barrier layer) are being polished back. Refer to Chaps. 15 and 16 for more information on these topics.

Erosion is an effect that occurs during CMP of Cu or W which results in the thinning of the ILD layer. That is, erosion is defined as the difference in the ILD thickness before and after the metal-CMP-step. The Cu (or W) film is also thinned, but its final thickness is reduced by the sum of the dishing plus erosion that has occurred. Erosion occurs because there is also a non-zero removal rate during the over-polish of the metal.

The extent of erosion depends heavily on the pattern-density (but not on metal linewidth, as is the case with dishing). Erosion is worst in regions of dense metal pattern (Fig. 8-99). This is so because when the polish begins the metal (e.g., Cu) covers the entire wafer. In regions of high Cu-pattern-density there is a lower

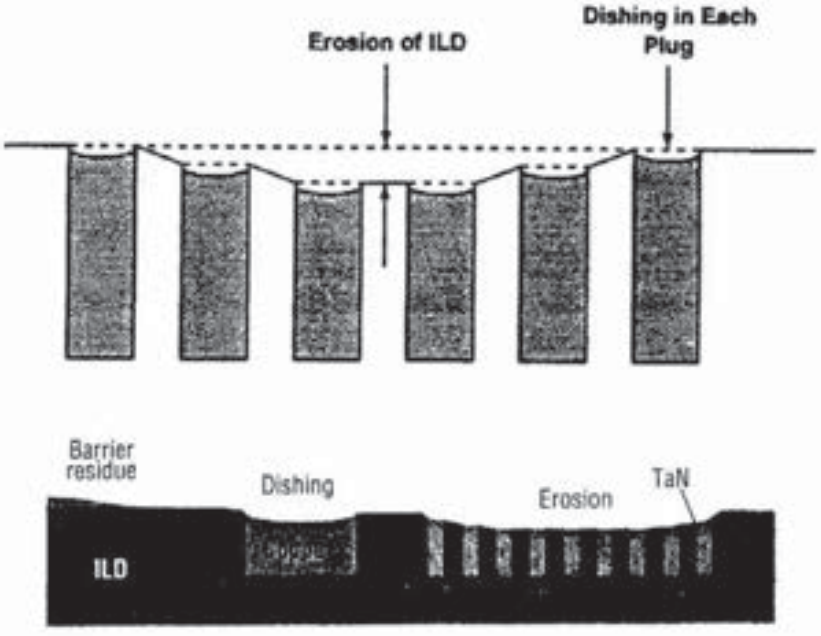


Fig. 8-99 Erosion of the ILD that occurs during over-polish of Cu during CMP.

aerial-density of Cu contacting the pad ("up-areas" of Cu). This is because the Cu in low areas (the "down-areas" of Cu) does not contact the pad. Thus, there is a higher pressure on the "up-areas" of Cu (and consequently higher polish-rate) than in regions where the Cu pattern-density is lower. Because of this difference in polish-rates, Cu is cleared in regions of high pattern-density sooner than in regions of lower pattern-density. Thus, the ILD in regions of high pattern-density receives a longer over-polish, adding to increase in erosion with pattern density.

8.12.2 Thickness Non-Uniformity Within a Wafer After CMP

Rigid polishing-pads provide good planarity within a die (chip), but poor thickness-uniformity across a wafer, for reasons explained here. If wafers were perfectly flat, this problem would not occur. A film on the surface of a flat wafer would be polished to a uniform thickness across the wafer (Fig. 8-100a). But not all wafers are uniformly flat. For example, if a wafer is tapered and the pad is rigid, the film on the thicker end of the wafer will be polished more than on the thinner end (Fig. 8-100b). If the wafer thickness undulates (with some regions thicker than others), a film covering regions where the wafer is thicker will be polished more (and will end up thinner) than a film on the thin parts (Fig. 8-100c). Softer (i.e., more compliant) pads produce improved uniformity across-the-wafer thickness, but at the price of degraded planarity. As discussed earlier,

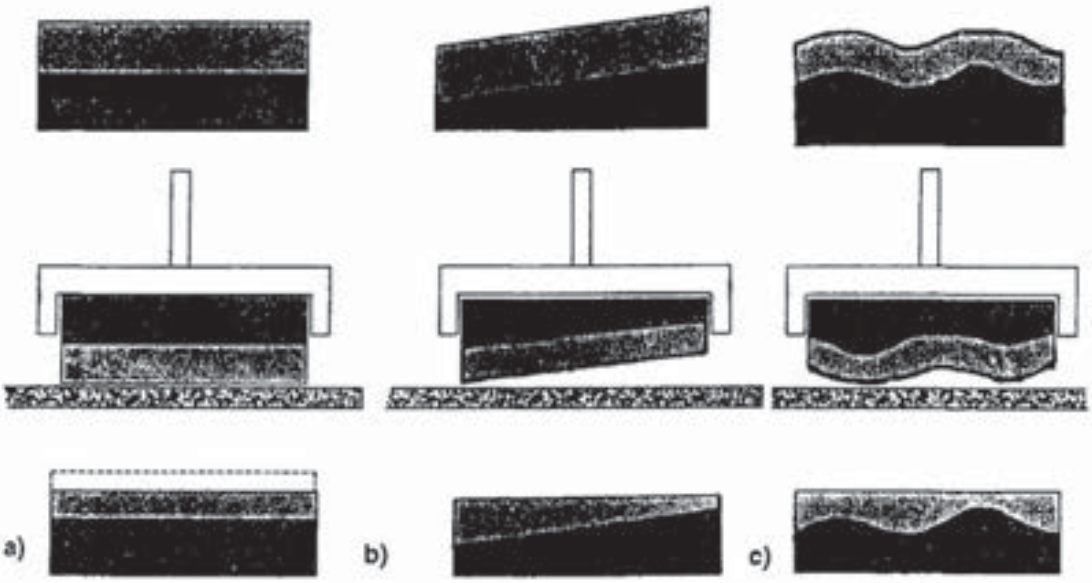


Fig. 8-100 If a rigid pad is used for CMP: a) A film on a flat-wafer is uniformly polished; b) A film on a tapered-wafer is polished thinner over the thicker end of the wafer; (c) For wafers with variations in wafer-thickness across the wafer, the film is polished thinner over the thicker wafer regions.

one solution is to use a composite pad (a rigid-pad backed by a softer one). Another approach is to use a linear polishing tool with a flexible (e.g., air) bearing beneath the polishing belt. This allows the belt to conform better to the wafer-surface, and reportedly produces improved across-the-wafer thickness-uniformity (and significantly less dishing).

8.12.3 Economic Considerations: Throughput and Cost of Ownership

While CMP is an enabling process for deep-submicron IC-fabrication, it is also a relatively expensive one. The cost of a single CMP-pass in 1999 (without a cleaning process) ranged from \$5 to \$8, depending on the material being polished. The post-CMP cleaning-step adds about another \$1 to \$2 per wafer. The main contributors to the cost of CMP include the polishing (and cleaning) tools, the slurry, the test- and monitor-wafers, and the need to do off-line measurements. The CMP-facilities, labor-costs, utilities, DI-water, pads, and waste-disposal are additional expenses. The 2002-target for the cost of a CMP process on 300-mm wafers is \$5.56/wafer/pass.[101]

The price of a fully-integrated polishing-tool in 1999 ranged from \$2–2.5 million, with the cleaning-module costing about \$500,000. Such tools are capable of polishing 50-60 200-mm wafers-per-hour. Slurries are also expensive. Oxide-slurries cost about \$10/gallon, tungsten-slurries about \$30/gallon, and Cu-slurries from \$40–50/gallon. One gallon is consumed for about every ten wafers.

Test- and monitor-wafers are needed not only during the set-up of the process, but also to check tool-performance for every lot. Typically, the removal-rate and the within-wafer-non-uniformity are monitored for each lot to ensure the tool is operating within spec. Some IC manufacturers have reported that up to two hours per shift are needed to process and measure the monitor-wafers and product test-wafers off-line, and that this reduces the CMP-equipment-capacity by nearly 50%. This is the reason that *in situ* measuring techniques are so attractive.

When the *cost-of-ownership* (CoO) of various CMP-tools are being compared, the three parameters that have the most significant impact on this cost are:

1. Throughput
2. Mean-Time-Between-Failure (MTBF)
3. Availability

The *throughput* of a CMP-tool is perhaps the most important parameter that contributes to the cost-of-ownership. It is defined as the quantity of wafers that can be completely polished within a given period of time (e.g., wafers-per-hour). A throughput of 50-60 wafers-per-hour is considered to be a good throughput

range. This parameter is highly dependent on the number of polishing platens and carrier heads. The 2002 performance-target for CMP-tool-throughput is 75 wafers-per-hour-per-tool.[102]

The *mean-time-between-failure* (*MTBF*) is a measure of a tool's reliability. It is defined as the mean-value of the number of hours that a specific tool can run continuously, non-stop, without breaking down. Most CMP tool-manufacturers seek to design their tools with an MTBF of at least 400 hours.

The *availability* of a CMP tool is the actual time the tool can run product wafers (i.e., the time when a tool is being qualified or tested, or is shut down for preventative-maintenance is subtracted from the total-time a tool is in service). This parameter reflects the ease with which a tool can be serviced or maintained

It should be added as a final note that the test-time is not always taken into account when calculating CoO. However, the amount of time required to qualify a CMP-tool for processing and the *mean-time-to-test* (*MTTT*), can also have a noticeable impact on CoO. If one assumes that a CMP-tool needs to be qualified once during each shift (i.e., three times a day), and that each qualification test (i.e., off-line thickness-measurement) takes about 30 minutes, this means that the MTTT is 1.5 hours per day, or 10.5 hours per week. This 10.5 hours could be considered lost CMP-processing-time. The shorter is this time, the higher the CMP throughput-number, which would lower the CoO.

REFERENCES

1. L. Rothman, *J. Electrochem. Soc.*, **127**, 2216 (1980).
2. C. Jang *et al., Proc. 4th Intl. IEEE VMIC Conf.*, Santa Clara CA, p. 357, 1987.
3. S. Morimoto *et al.*, "Characterization of the Chemical-Mechanical Polishing of Intermetal Dielectric Films," *J. Electrochem. Soc.* 1993, p. 449.
4. J.M. Steigerwald, S.P. Murarka, and R.J. Gutmann, *Chemical Mechanical Planarization of Microelectronic Materials*, Wiley & Sons, New York, 1996.
5. M.A. Fury, "The Early Days of CMP," *Solid State Technol.*, May 1997, p. 81.
6 B. Davari *et al*, "A New Planarization Technique, Using a Combination of RIE and Chemical Mechanical Polishing (CMP)," *Tech. Dig. IEDM*, 1989, p. 61.
7. R. DeJule, "CMP Grows in Sophistication" *Semicond. Internatl*, Nov. 1998, p. 56.
8. M. Tomozawa, "Oxide CMP Mechanisms," *Solid State Tech.*, July 1997, p. 169.
9. F. Preston, "Theory & Design of Plate Glass Polishing Machines," *J. Soc. Glass Tech.*, **11**, 214 (1927).
10. Paper TuB-A4, *Tech. Dig., Topical Meeting on the Science of Polishing*, Optical Soc. of America, 17 April 1984.
11. C.W. Liu *et al.* "Characterization of CMP Process Based on Nanoindentation Measurement of Dielectric Films," *J. Electrochem. Soc.*, **142**, p. 3098 (1995).

12. S.R. Runnels and L.M. Eyman, "Tribology Analysis of CMP," *J. Electrochem. Soc.*, **144**, p. 1698 (1994).
13. W.T. Tseng, "Re-examination of Pressure & Speed Dependence of Removal rates during CMP processes," *J. Electrochem.. Soc.*, **144**, p. L15 (1997).
14. B. Zhao and F.G. Shi (1999). "CMP: Threshold Pressure and Mechanism," *Electrochemical Solid State Letters* **2**, p. 145 (1999).
15. G.B. Shinn *et al.*, "Chemical-Mechanical Polish," in *Handbook of Semiconductor Manufacturing Technology,* Y. Nishi and R. Doering, Eds., Marcel Dekker, New York, 2000, p. 415.
16. "New Model Could Vastly Improve CMP Processes," *Semiconductor International,* May 1999, p. 21.
17. L. Cook, "Chemical Processes in Glass Polishing," *J. Non-Crystalline Solids*, **126**, 160 (1990).
18. W.-T. Tseng *et al.*, "CMP of Fluorinated Silicon Dioxide: Is it Necessary and Feasible?" *Solid State Technology*, February 1997, p. 64.
19. W.J. Schaffer and H.W. Fry, "BPSG Improves CMP Performance for Deep Submicron ICs," *Semiconductor International*, July 1996, p. 205.
20. T. Izumitani and S. Harada, "Polishing Mechanism of Optical Glasses," *Glass Technology,* **12**, 131 (1971).
21. *ibid. Ref. 17.*
22. J. Olsen and F. Moghadam, "Planarization Techniques," in *Handbook of Multilevel Metallization for Integrated Circuits,* Eds. S.R. Wilson, C.J. Tracy, and J.L. Freeman, Noyes Pubs., Park Ridge, NJ, 1993, Chap. 6, pp. 426-438.
23. W.J. Patrick *et al.*, "Application of CMP to the Fabrication of VLSI Circuit Interconnections," *J. Electrochem. Soc.*, **138**, June 1991, p. 1778.
24. P.A. Burke, "Semi-Empirical Modeling of SiO_2 CMP Planarization," *Proc. VMIC* 1991, p. 379.
25. J. Schlueter, "Controlling Contamination During CMP Processes," *Solid State Technology,* May 1996, p. S8.
26. F. Kaufman., "Chemical Mechanical Polishing for Fabricating W Metal Features as Chip Interconnects," *J. Electrochemical Society,* **138**, Nov.1993, p. 3460.
27. M. Pourbaix, *Atlas of Electrochemical Equilibria in Aqueous Solutions,"* Nat. Assoc. of Corrosion Engr., Houston TX (1974).
28. J. Gisler, "Enhancing CMP Performance, *Semicond. Internatl,* July 2000,p. 150.
29. C.C. Yu., *Advanced Metallization for ULSI Applications Proceedings*, Oct., 1993.
30. A.E. Braun, "CMP Battles Low-k Fundamentals," *Semiconductor International,* October 2000, p. 72.
31. K. Witt and L. Cook, "Getting an 'Edge' with CMP," *Semiconductor International,* October 2000, p. 75.
32. Courtesy of Rodel Inc.

33. Courtesy of Rodel Inc.
34. H. Nojo, *et al.*, "Slurry Engineering for Self-Stopping & Dishing-Free SiO_2 CMP," *Tech. Dig. IEDM*, 1996, p. 349.
35. A.E. Braun, "Slurries and Pads Face 2001 Challenges," *Semiconductor International*, November 1998, p. 65.
36. T. Hara *et al.*, "Direct CMP of Low-*k* Organic by MnO_2 Slurry," *Mtg. Abs. Electrochem. Soc.*, Fall 1999, Abs. 1269.
37. J. McCann and C. Smith, "Bulk Slurry Distribution for Integrated Manufacture," *Semiconductor International*, November 1996, p. 94.
38. G. Basim *et al.*, "Effect of Particle Size of Chemical Mechanical Polishing Slurries for Enhanced Polishing with Minimal Defects," *J. Electrochem. Soc.*, September 2000, p. 3523.
39. I. Ali *et al.*, "Chemical-Mechanical Polishing of Interlevel Dielectric: A Review," *Solid State Technology*, October 1994, p. 63.
40. J.H. Golden *et al.*, "Evaluating and Treating CMP Wastewater," *Semiconductor International*, October 2000, p. 85.
41. H.-J. Kim *et al.*, "Physical and Chemical Characterization of Re-Used Oxide CMP Slurry," *Mtg. Abs. Electrochem.. Soc.*, Fall 2000, Abs. 477.
42. *ibid. Ref. 4.*
43. Courtesy of Rodel Inc.
44. Courtesy of Rodel Inc.
45. A.E. Braun, "Slurries and Pads Face 2001 Challenges," *Semiconductor International*, Nov. 1998, p. 65.
46. M.R. Oliver *et al.*, "CMP Pad Surface Roughness and CMP Removal Rate," *Mtg. Abs. Electrochem. Soc.*, Fall 2000, Abs. 487.
47. I. Ali and S.R. Roy, "Pad Conditioning in Interlayer Dielectric CMP," *Solid State Technology*, June 1997, p. 185
48. C.-Y. Chen *et al.*, "Operational Aspects of Chemical Mechanical Polishing: Polishing Pad Profile Optimization," *J. Electrochem. Soc.* October 2000, p. 3922.
49. Courtesy of 3M Company.
50. Courtesy of 3M Company.
51. Courtesy of 3M Company.
52. W.L. Smith *et al.*, "Film Thickness Measurements for Chemical Mechanical Planarization," *Solid State Technology*, January 1996, p. 77.
53. A.S. Lawing, "Pattern Dependent Polish Rate Behavior in Oxide Chemical Mechanical Polishing," *Mtg. Abs. Electrochem. Soc.*, Fall 2000, Abs. 493.
54. E. Worthington, "New CMP Architecture Addresses Key Process Issues," *Solid State Technology*, January 1996, p. 61.
55. R. DeJule, "Advances in CMP," *Semiconductor International*, November 1996, p. 88. 56. Courtesy of Applied Materials

57. H.E. Litvak and H.-M. Tzeng, "Implementing Real-Time Endpoint Control in CMP," *Semiconductor International,* July 1996, p. 259.
58. H.E. Litvak, *U.S. Patent No. 5,499,733.* Issued March 19, 1996.
59. H. Litvak, *Proc of the SEMI Taiwan CMP Seminar* Sept. 1995, Taipei, p. 123.
60. *ibid.* ref. 25.
61. J. Skrovan, K. Robinson. "Planarization slurry including a dispersant and method of using same." *U.S. Patent 5,827, 781.* Oct. 27, 1998.
62. F. Preston, "Theory & Design of Plate Glass Polishing Machines" *J. Soc. Glass Tech.,* **11**, 214 (1927).
63. V.B. Menon, R.P. Donovan, in W. Kern Ed., *Handbook of Wafer Cleaning Technology,* Noyes Publications, Park Ridge, N.J., 1993.
64. K. Mikhaylich *et al.,* "Evaluation of Contact and Non-Contact Cleaning for CMP Applications," *Mtg. Abs. Electrochem. Soc.,* Fall 1999, Abs. 1268.
65 D. Hymes *et al.,* "Brush Scrubbing Emerges as a Future Wafer-Cleaning Technology," *Solid State Technology,* July 1997, p. 209.
66. T.L. Myers, "Post-Tungsten CMP Cleaning," *Solid State Tech.* Oct. 1995, p. 59.
67. J.M. de Larios *et al.,* "Post-CMP Cleaning of Oxide and Tungsten Applications," *Semiconductor International,* May 1996, p. 121.
68. D. Hymes *et al.,* "The Challenges of the Copper CMP Clean," *Semiconductor International,* June 1998, p. 117.
69. K. Osseo-Asare, K.K. Mishra, "Solution Chemical Constraints in the CMP of Copper: Aqueous Stability Diagrams for the Cu-H_2O and Cu-NH_3-H_2O Systems," *J. Electron. Materials,* **25**, p. 1599 (1997).
70.Y. Feng, *et al.,* "Corrosion Protection of Copper by a Self-Assembled Monolayer of Alkenathiol," *J. Electrochem. Soc.*, **144**, p. 55 (1997).
71. A.A. Busnaina and F. Dai, "Megasonic Cleaning," *Semiconductor International.* August 1997, p. 85.
72. A. Busnaina *et al.,* "Single Wafer Non-Contact Post-CMP cleaning Using DI Water and Dilute Chemistry," *Mtg. Abs. Electrochem. Soc.,* Fall 1999, Abs. 1267.
73. M. Elissa *et al.,* "Post Copper CMP Cleaning: A Non-Contact Method," *Mtg. Abs. Electrochem. Soc.,* Fall 1999, Abs. 1270.
74. I. Ali *et al.,* "Investigating the Effect of Secondary Platen Pressure on Post-CMP Cleaning," *Microcontamination* **12**(10), p. 45, 1994.
75. R. Davenport and B. Brown, "Integrated CMP Clean Methods," *European Semiconductor,* July 1999, p. 36.
76. Courtesy of Ebarra.
77. Y. Zhang *et al.,* "Wafer Dimensional Analysis for Chemical Mechanical Planarization," *Solid State Technology,* July 1997, p. 179.
78. J. Reynolds *et al.,* "Defect Inspection Technologies to Meet the Challenges of Advanced CMP Processes," *Solid State Technology,* March 1998, p. 39.

79. Courtesy of KLA-Tencor Inc.
80. *ibid., Ref. 55.*
81. *ibid., Ref. 39.*
82. W. O'Mara, "The Inevitable Shakeout in the CMP Equipment Market," *Solid State Technology*, May 1998, p. 84.
83. F. Kaufman *et al.,* "Chemical Mechanical Polishing for Fabricating Patterned W Metal Features as Chip Interconnects," *J. Electrochemical Society,* Vol. **138**, November 1993, p. 3460.
84. K. Wijekoon *et al.,* "Tungsten CMP Process Developed," *Solid State Technology,* April 1998, p. 53.
85. I. Kim *et al.,* "Optimized Process Developed for Tungsten CMP," *Semiconductor International,* November 1996, p. 119.
86. S. Roy, "Advanced Metal & Intercon. Syst. for ULSI Applications," 1995, p. 733.
87. *ibid. Ref. 29.*
88. J.-F. Wang "CMP of Cu Multilevel Interconnects," *Procs. VMIC* 1995, p. 505.
89. R. DeJule, "CMP Challenges Below Quarter Micron," *Semiconductor International,* Nov. 1997, p. 54.
90. D.R. Evans, "Electrochemical Interaction Between Copper and Barrier Materials During CMP," *Mtg. Abs. Electrochem. Soc. Mtg.*, Fall 1996, Abs. No. 504,
91. A. Beverina *et al.,* "Copper Photocrrosion Phenomenon during Post-CMP Cleaning," *Electrochem. Soc. Letts.,* March 2000, p. 156.
92. A.R. Seturaman *et al.,* "Copper vs Aluminum: A Planarization Perspective," *Semiconductor International,* June 1996, p. 177.
93. J. Nuetzel *et al.,* "Conventional and Slurry Free CMP of Polymer Layers for Gbit DRAM," *Mtg. Abs. Electrochem. Soc.,* Fall 1999, Abs. 1232.
94. W.-T. Tseng, "CMP of Fluorinated Silicon Dioxide: Is it Necessary and Feasible?" *Solid State Technology*, February, 1997, p. 61.
95. *ibid. Ref. 36.*
100. S.J. Fang *et al.,* "Optimization of the CMP Process for Pre-Metal Dielectrics," *J. Electrochem. Soc.,* February 2000, p. 682.
101. T. Marbeiter, *et al.,* "An Update: Transition to 300-mm CMP," *Semiconductor International,* November 1998, p. 76.
102. *ibid. Ref. 101.*
103. M. Nandakumar *et al.,* "Shallow Trench Isolation for Advanced ULSI CMOS Technologies," *Tech. Dig. IEDM,* 1998, p. 134.
104. L. Peters, "Choices and Challenges for Shallow Trench Isolation." *Semiconductor International,* April 1999, p. 71.
105. G. Badenes *et al.,* "A New Dummy-Free Shallow Trench Isolation Concept for Mixed-Signal Applications," *J. Electrochem. Soc.,* **147**, Oct. 2000, p. 3827.

Chapter 9

SHALLOW TRENCH ISOLATION (STI)

Despite advances made to decrease bird's beak and channel-stop encroachment in LOCOS isolation (by poly-buffered LOCOS and other approaches), such techniques eventually become inadequate for deep-submicron IC technologies. In particular, the scalability of LOCOS for sub-half-micron CMOS technologies comes into question. The limitations of LOCOS include: the penetration of the LOCOS bird's beak into the device active regions; non-planarity; LOCOS field-oxide-thinning at narrow isolation spacings; and stress-induced silicon defects. (LOCOS-based isolation technologies are described in more detail in Vols. 1 [2nd Ed], 2, and 3 of *Silicon Processing for the VLSI Era.*) The successor to LOCOS in CMOS is *shallow trench isolation* (STI). In STI a relatively shallow trench is first etched into the Si substrate (0.3-0.5-μm deep), which is then refilled with an insulator material (following a short thermal-oxidation growth step that forms a thin film of SiO_2 on the trench walls to control the Si-SiO_2 interface quality). The surface is planarized by CMP after refill to complete the isolation structure. Figure 9-1a depicts a shallow-trench isolation (STI) structure in CMOS.

Since the bird's beak is entirely eliminated in STI and the redistribution of channel-stop-dopant is reduced, smaller isolation spacing is also possible than with LOCOS. In addition, the field oxide in STI is fully recessed, offering the potential of a completely planar surface after the isolation structure formation. (Note that while CMP can also be used to reduce topography resulting from the LOCOS process, the oxide encroachment with LOCOS is not eliminated by such CMP.) Furthermore, the problem of field-oxide-thinning in narrow isolation spaces does not occur, and the threshold voltage is constant as a function of channel width (because of the elimination of the narrow-channel effect found in LOCOS - see Chap. 9, *Silicon Processing for the VLSI Era*, Vol. 3). Also, MOSFETs made with fully-recessed oxide isolation exhibit larger g_m and smaller S_t values than those using semi-recessed LOCOS. Finally, the fully-recessed STI structures have trenches whose lower corners, though rounded, have

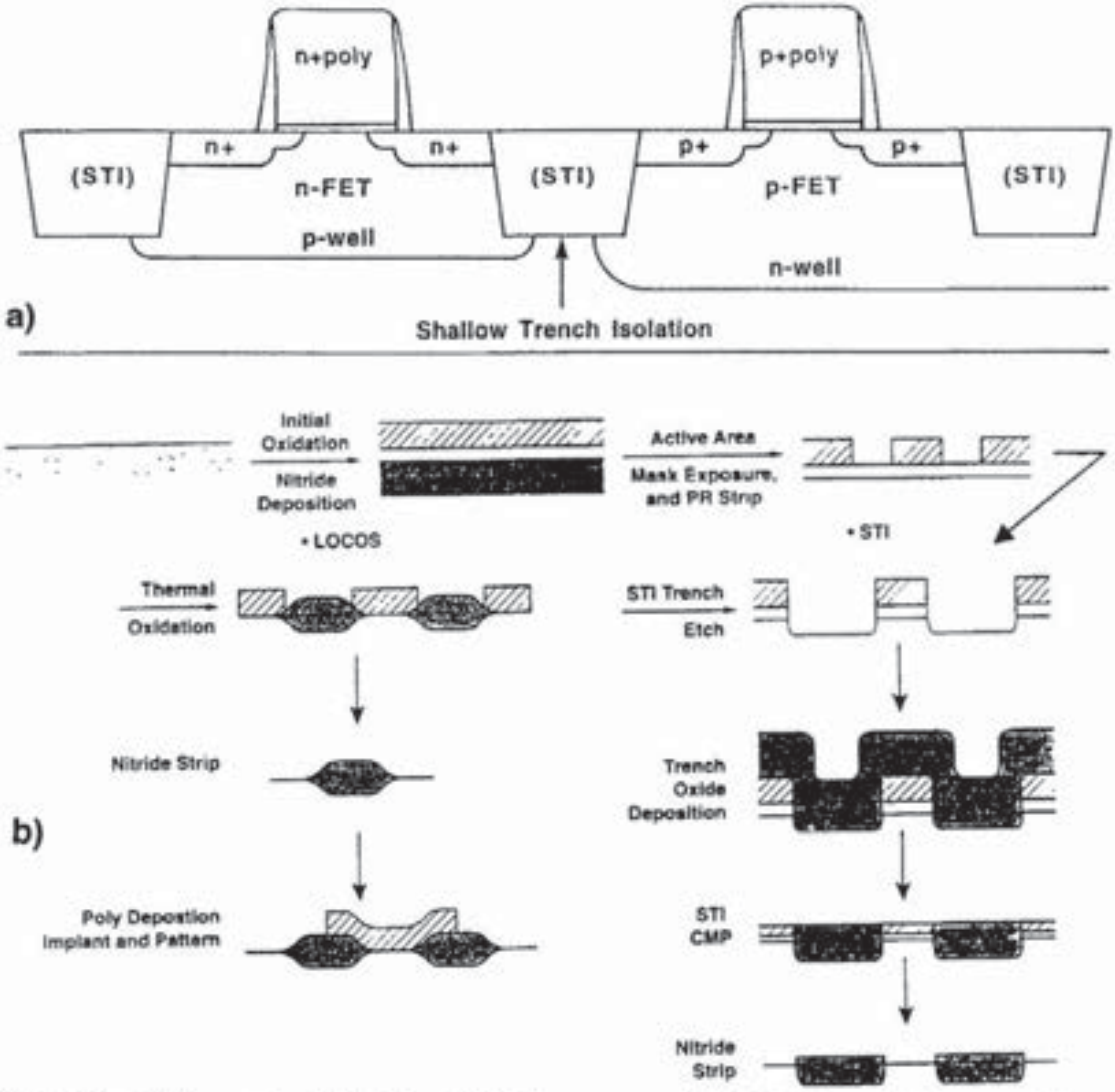


Fig. 9-1 a) Schematic of STI in a CMOS structure. b) LOCOS vs. STI process flow, indicating that STI requires more steps.

a smaller curvature than exists with semi-recessed LOCOS field oxides. These trench corners provide improved isolation efficiency due to the so-called "corner effect."[1] That is, the field lines at the bottom corners of the trench spread out at the vicinity of the corner, leading to a higher threshold voltage at the corners. Until the silicon surface at these corners is inverted, isolation is maintained.

Despite the impressive advantages of shallow trench isolation listed above, a variety of new difficulties associated with the fabrication and process integration of STI structures arise. For one thing, the STI process involves a larger number of steps than does LOCOS isolation process (Fig. 9-1b). Yet, the overall benefits of STI over LOCOS provided the impetus for the industry to find ways to

overcome these challenges. STI began to find use in mainstream IC technology when it was integrated into 166 MHz Pentium microprocessors (which used a 0.35-μm process) and the PowerPC 603e chip (which was fabricated with a 0.45-μm process). In this chapter we will describe the details of the STI process fabrication sequence, the process integration challenges that have been overcome, and those that it faces going forward.

9.1 EARLY SHALLOW-TRENCH-ISOLATION STRUCTURES

The earliest example of shallow-trench-and-refill is the so-called *buried-oxide* (or BOX) isolation technology. The original BOX technique (BOX-I) was introduced in 1981,[2] and an improved version was presented in 1983 (BOX-II).[3] In the latter, shallow trenches (0.5-0.8 μm deep) were etched with an orientation-dependent silicon-etch to provide an inclination of the trench sidewalls of ~60°. Next a channel-stop-implant step was performed. (The sloped sidewalls of the trenches ensure that they receive sufficient doping during the channel-stop implant step to prevent weak inversion of the silicon-sidewall surface. Thus, subthreshold leakage from source-to-drain along this parasitic conduction path is suppressed.) A thin thermal oxide (20-50-nm-thick) was then grown to passivate the trench sidewalls. Next, CVD oxide was deposited to fill the trenches. This layer was etched back, leaving oxide only in the recesses. Ideally, the top of the oxide should be level with the original silicon surface. Etchback was performed by eroding photoresist and SiO_2 at the same rate in a plasma (see Chap. 4, Vol. 2). In the improved process, two layers of resist were applied (to try to produce a globally planarized surface prior to etchback). Active regions are those protected from etch when the trenches are formed.

These early BOX processes, however, exhibited several drawbacks, including:

1. Void formation occurred if the CVD oxide was deposited into 0.5-μm-deep trenches that are narrower than about 2-μm.

2. Inversion of the silicon at the sidewalls of p-type active regions could still occur, especially if vertical trench sidewalls were used to reduce the isolation spacing. This problem was made worse by boron segregation into the thin, sidewall-passivating thermal oxide. This effect lowers the boron concentration at the sidewall surface, as will be discussed in following section.

3. Despite complex planarization schemes, the early BOX processes were unable to yield adequately planar surfaces. (In particular, global planarity was poor.)

4. The BOX isolation structures could negatively impact the device operating behavior.[4,5,6,7] That is, in a conventional LOCOS-isolated MOSFET, a polysilicon gate that passes across the device will curve upward and away from the device once it reaches the isolation oxide, as shown in Fig. 9-2a. (This figure represents the cross section of the region under the gate of the MOSFET, where the source and drain of the device are situated above and behind Fig. 9-2, which lies in the plane of the paper.) For this particular configuration of the polysilicon gate, the portion of the gate poly over the LOCOS region will have little effect upon the device behavior. Consequently, a simple 2-D simulation will give an adequate description of the behavior of the device, where the 2-D region is taken in the plane perpendicular to Fig. 9-2 and along the axis of the device, so that it includes the source and drain regions.

In the ideal case of a shallow-trench-isolated MOSFET, the gate poly runs flat across the isolation, as shown in Fig. 9-2b. Consequently, the part of the gate poly over the field region creates more of a two-dimensional fringing electric-field on the corner and sidewall of the silicon mesa than in the LOCOS case of Fig. 9-2a. This establishes a *corner parasitic-MOSFET* in parallel with the main device.[8] The parasitic device turns-on at gate voltages lower than those required by the main channel (because the same voltage value occurring at the corner produces a larger electric field at that point, inducing a higher concentration of

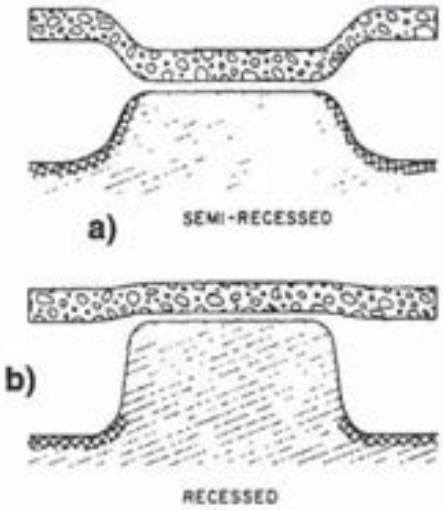


Fig. 9-2 a) In LOCOS isolation structures, the field oxide rises above the substrate. b) In trench isolation structures, the field oxide is flat with respect to the substrate (unless the trench oxide is overetched; in that case, a downward step exists at the edge of the active regions – see, for example Fig. 9-3).

inversion-layer charge than the same voltage induces in the nominal channel region). This causes a "hump" or "kink" in the drain-current versus gate-voltage subthreshold curves (see Fig. 9-3). Thus, the parasitic device increases the subthreshold leakage-current of the active device.

This effect becomes even stronger if the isolation region is partly etched below the top surface of the silicon mesa (i.e., the surface of the active region of a MOSFET IC). In that case, the polysilicon gate wraps around the corner of the mesa as shown in Fig. 9-3. Now the regions of the channel near the edges of the device will behave much differently than the region at the center of the mesa. Consequently, 3-D modeling is required to accurately simulate the device characteristics. Such simulations verify that the corner regions have a lower threshold voltage than do the center regions of the device.

5. Thinning of the gate oxide at the trench edge is another concern. Thinned gate oxide not only degrades the gate-oxide-reliability, but also worsens the parasitic device problem (since a thinner gate oxide increases the electric-field-strength at the trench corner).

6. Problems arose as result of the etchback step. That is, to ensure complete removal of the CVD SiO_2 from the active areas, the field oxide must be overetched by about 100 nm. The field oxide surface is then lower than active-area surface, causing a portion of the active-area sidewall to become exposed. This exacerbates the parasitic sidewall conduction described above and also enhances the inverse narrow-width effect (see Chap. 6, *Silicon Processing for the VLSI Era* - Vol. 3).

The combination of the above problems (especially the complex processing required to produce adequately planar surfaces), kept BOX-isolation from displacing LOCOS-type isolation-structures. However, the benefits offered by shallow-trench isolation are so attractive that work continued in an effort to overcome the above problems. Several enabling technologies were developed to overcome the drawbacks of the early shallow-trench isolation processes. These included: *dry-etch processes* that allowed trenches with controlled sidewall slopes to be produced; *deposition techniques* that permitted the trenches to be filled without the formation of voids; and *chemical-mechanical polishing* (CMP) methods that could provide global planarization of the wafer surface at the conclusion of the STI process. We will describe these technologies in more detail in the next section, emphasizing how their use enabled the implementation of STI. Figure 9-4 shows an SEM of an STI structure.

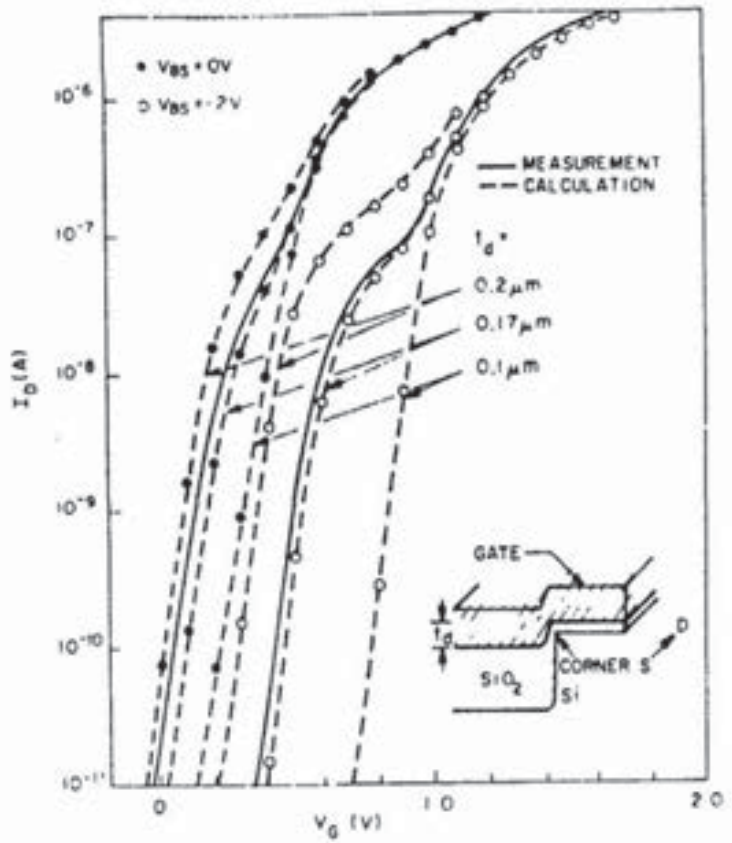


Fig. 9-3 Measured and simulated drain current versus gate potential for devices having a downward step in the field oxide (t_d).[9] (© 1981 IEEE)

9.2 STI-ENABLING PROCESSES

Progress in overcoming the four major problems of early-STI isolation structures is summarized as follows:

1. The problem of void formation when refilling narrow trenches with CVD-oxide was ameliorated in two ways: **1**) dry-etch processes which could etch the trenches with a controlled slope (and rounded-bottom trench corners) were developed to replace the wet-etch techniques used to etch the trenches in early-STI structures; and **2**) various so-called "gap-filling" CVD techniques were developed that could fill the trenches with CVD-oxides without creation of voids. Such gap-filling-CVD processes include: *LPCVD TEOS*; *atmospheric-pressure CVD of ozone-TEOS*; and *high-density-plasma CVD* (*HDP-CVD*). Note that the film stress of the trench-fill dielectrics must also be smaller (<200 MPa) than in intermetal dielectric films, due to the use of thicker films.

2. The introduction of *chemical-mechanical polishing* (*CMP*) allowed the

oxide that is deposited into the trench to be polished back, leaving a planar surface over the wafer. The CMP process was developed in the late 1980s at IBM, and was introduced to the rest of the IC industry in the early 1990's. Details of the CMP process are provided in Chap. 8.

3. As described above, the subthreshold characteristics of narrow MOSFETs (e.g., $\leq$ 0.5-μm-wide) isolated by refilled shallow-trench structures are sensitive to the shape of the upper corners of the trenches.[10] That is, the edges of the silicon mesa at the upper corners are turned-on by the portion of the gate that overlaps such corner regions (see also Sect. 5.2.2 in *Silicon Processing for the VLSI Era*, Vol. 3). This edge-effect becomes even stronger if the isolation region is partly eroded away, in which case the gate wraps around the corner of the device, as shown in Fig. 9-3. The optimum condition for the shallow-trench-isolation structure is thus a gate that runs flat across the isolation, requiring a planar surface prior to poly deposition. A process utilizing RIE-etchback followed by a chemical-mechanical-polishing step offers the ability to produce a highly planar surface after etchback. The top surface exhibits virtually no step. However, single-wafer processing with *in-situ* control of the CMP process is likely to be needed to obtain the exact surface required.

9.3 SHALLOW TRENCH ISOLATION FOR CMOS

Trench isolation has also been explored for twin-tub CMOS as a means of implementing isolation structures, primarily as a means for achieving smaller interwell (n^+-to-p^+) isolation spacing. (Other advantages of trench isolation over LOCOS were described earlier.) A trench at the n-well-to-p-well interface prevents dopant interdiffusion where the two wells join. Thus, dopant compensa-

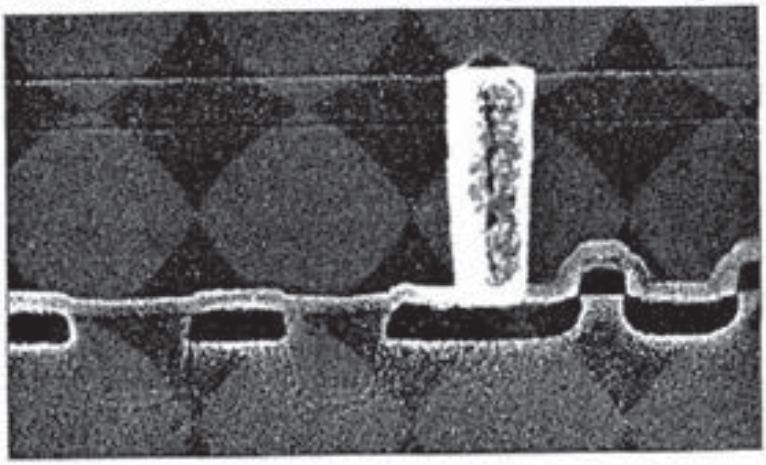


Fig. 9-4 SEM of a shallow-trench isolation structure.

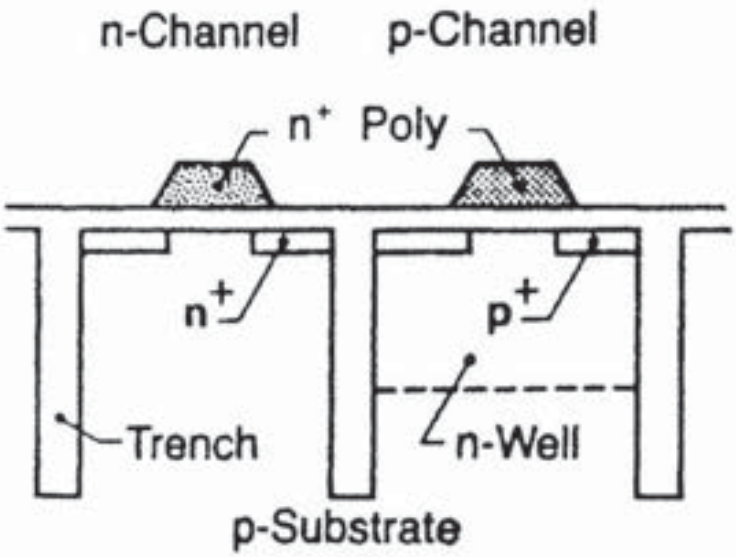


Fig. 9-5 Schematic of an *n*-well CMOS with deep trench isolation.[11] (© 1986 IEEE).

tion (which lowers the surface dopant concentration at the well borders) is eliminated. This permits larger parasitic-MOSFET-threshold-voltage values (V_{TFiso}) to be maintained at smaller isolation spacings. Shallow and deep trenches of submicron widths have been used to reduce p^+-to-n^+ spacings.

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. 9-5). However, since fabrication of deep-trench-isolation for CMOS is substantially more complex than shallow-trench-isolation, and latch-up measurements have not fulfilled expectations, this approach was not widely implemented for such applications. Shallow-trench-isolation structures, however, became the standard isolation structure for CMOS at about the 0.25-μm node.

It should be noted, however, that while their use permits smaller n^+-to-p^+ spacing, the presence of trench structures also gives rise to additional leakage paths not present in LOCOS-CMOS-isolation structures. These new paths can degrade the isolation between the active devices and the well. In some cases these paths can also increase the subthreshold leakage-current within *active* devices. Generally, these new leakage phenomena involve the formation of an inversion channel along the vertical trench-sidewall (see Fig. 9-6a).

The sidewall inversion is caused by the *horizontal* parasitic MOS device, in which the positively biased *n*-well acts as the gate electrode, and the trench dielectric serves as the MOS gate oxide (with a thickness equal to the trench width). If the *n*-well is biased at $+V_{DD}$, and the trench is narrow (e.g., ≤ 1-μm), inversion along the sidewall of the trench between the source and drain of the

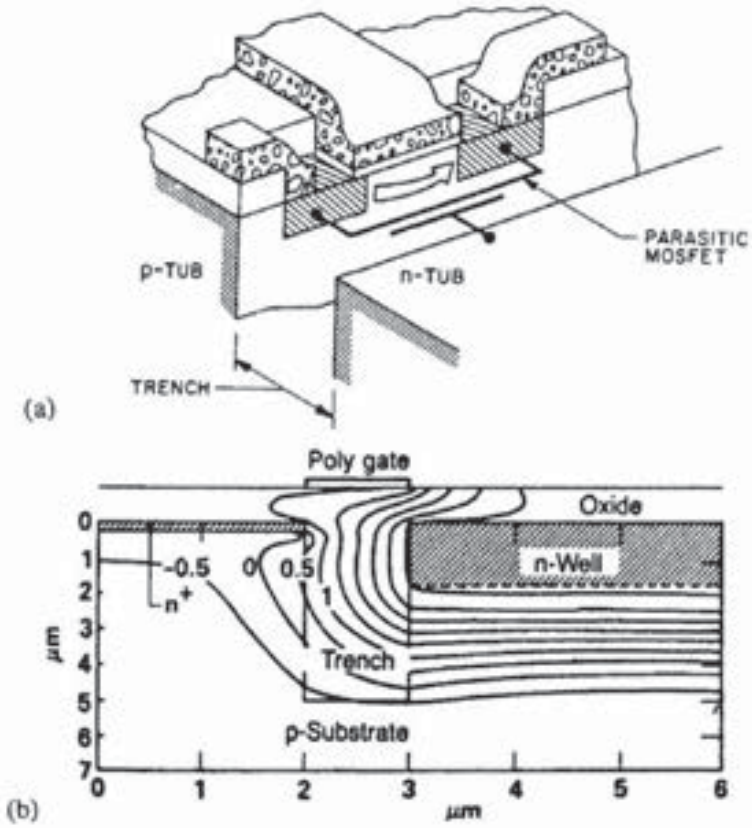


Fig. 9-6 a) A parasitic-sidewall-MOSFET with the *n*-well as the gate and the trench serving as the gate oxide.[12] From S.M. Sze Ed., *High-Speed Semiconductor Devices*, Copyright 1990 by John Wiley & Sons. Reprinted with permission. b) Cross-sectional view of a simulation of the 2-D equipotential contours in an *n*-well CMOS with trench isolation. The simulation is done at $Q_{fc} = 5x10^{10}$ cm^{-2}, $N_A = 9x10^{14}$ cm^{-3}, $V_{well} = 3V$, $V_{GS} = 0V$, $V_{sub} = -1V$.[13] (© 1983 IEEE)

MOSFET can easily occur. As seen in Fig. 9-6a, this is the trench sidewall that is outside of, but facing the *n*-well. In Fig. 9-6b, the simulated 2-D potential contours in a trench-isolated CMOS structure are shown. Note that the potential on the trench sidewall on the *p*-substrate side is > 0.5V, indicating that it is inverted.

It would appear that to take maximum advantage of such trench structures for reducing isolation spacing, the active devices should be placed right at the well edge (as shown in Fig. 9-6a). However, if an active NMOS-device abuts an inverted trench sidewall, a conducting path will exist between the source and drain of the active NMOS device. This path will exist regardless of the bias on the "real" gate of the active NMOS device, giving rise to an increase in the subthreshold leakage-current of the active device. In practice, to avoid this problem, active devices are usually moved back somewhat from the well edge.

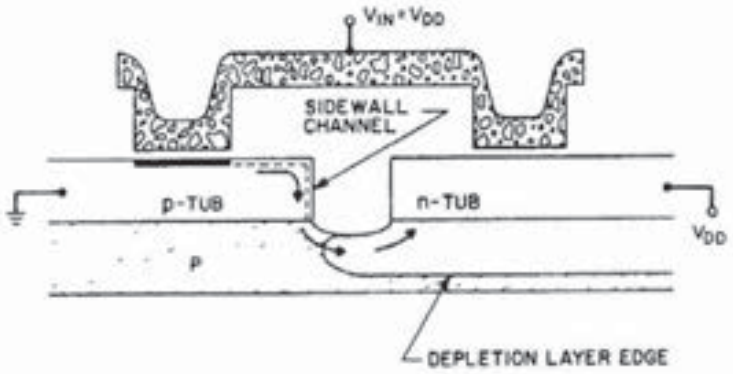


Fig. 9-7 A trench-isolated *p*-well and *n*-well with a parasitic vertical-sidewall-MOSFET (with the *n*-well as the gate) and a parasitic-horizontal-MOSFET (with the V_{in} poly-line as the gate).[14] From S.M. Sze Ed., *High-Speed Semiconductor Devices*, Copyright 1990 by John Wiley & Sons. Reprinted with permission.

Even if the NMOS device is moved back from the trench edge, the trench sidewall can still become inverted by a positively-biased *n*-well. If this happens and the poly over the field-region also inverts the top surface of the *p*-well, it is possible for a leakage path from the NMOS-source to the *n*-well to be established, as shown in Fig. 9-7.

Assuming that the leakage path of Fig. 9-7 is somehow suppressed (e.g., by implanting channel-stop-dopants into the bottom of the trench), the inverted trench sidewall can still produce other, unexpected leakage paths. One such dc leakage path between V_{DD} (e.g., 5V) and V_{SS} (e.g., 0V) is illustrated in Fig. 9-8 for the case of two series-connected CMOS inverters which suffer from inadequate CMOS isolation.[15] This path arises from leakage along both the inverted surface of the field-regions and the trench-sidewall (which, in this example, is offset from the active device edge). Here V_{in1} to the first inverter is at logic 1 (i.e., $V_{in1} = 5V$), causing the NMOS transistor of the first inverter to be turned *ON*. The output of the first inverter V_{out1} (and thus the input to the second inverter V_{in2}) is at logic 0. This causes the output of the second inverter V_{out2} to be at logic 1, and causes the PMOS device of the second inverter to be turned *ON*. The black highly conductive paths in Fig. 9-8 from V_{DD} to the edge of the n^+-region in the second inverter and from V_{SS} to the p^+-region in the first inverter are connected by a less-well-conducting parasitic conductive path. This path consists of an inverted region that runs under the poly of gate 1 in the *p*-well, then along the inverted trench sidewall, then under the poly of gate 2 (note that for clarity, in Fig. 9-8 the path under the poly of gates 1 and 2 is drawn as running adjacent to the poly, whereas the path actually exists beneath the poly).

More severe trench-sidewall-inversion problems are exhibited in p-wells than in n-wells. This is due to two interrelated phenomena. The first involves the dielectric (insulator) material used to fill the trench (typically SiO_2), and the resultant Si/SiO_2 interface at the trench-sidewall. The interface fixed or trapped charge that exists at the Si/SiO_2 interface is positive, which tends to lower the inversion-threshold of p-channel regions (in this case the p-well). To minimize this interface charge, a thin, good-quality thermal-oxide-layer must be grown inside the trench. This oxide growth process, however, gives rise to the second problem. That is, during thermal oxidation, some of the boron dopant in the p-well segregates into the oxide. Thus, the B surface-concentration at the interface is reduced, further decreasing the inversion-threshold. As a result, inversion of p-well trench-sidewalls occurs at lower threshold-voltages than do those of an n-well for comparable well-doping concentrations. In 0.25-μm (and smaller) CMOS technologies, this last problem is circumvented by forming the well-regions with an implant after the STI-structure is completed. A high-energy-implant process is used to implant well-dopants through the STI-oxide (and in

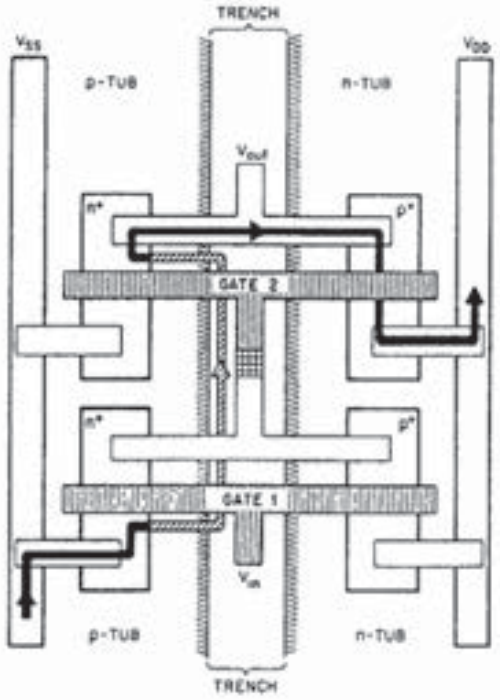


Fig. 9-8 A dc leakage path through two inverter stages in series. The solid black path is a good conducting route and the shaded path is a parasitic route. The parasitic consists of a field oxide corner under Gate-1 (cf. Fig. 9-7), a sidewall path along the trench (cf. Fig. 9-6a), and a corner path under the V_{out} strap.[15] From S.M. Sze Ed., *High-Speed Semiconductor Devices*, Copyright 1990 John Wiley & Sons. Reprinted by permission.

the active regions). Thus, the boron-segregation problem that would occur during the trench-liner-oxidation process is avoided.

To counteract intra-well and inter-well leakage in trench-isolated-CMOS several steps are necessary. First, the surface-leakage-path must be suppressed by use of sufficient channel-stop-doping between the active-device and well-edge (as is also done in LOCOS-isolated-CMOS). Second, to avoid subthreshold leakage in single active devices, the devices are moved back somewhat from trench edge. Third, inversion of the trench sidewall should be prevented. One way to do this is to increase the trench width. However, this will increase the isolation spacing (as will moving the active devices away from the active device to the trench edge).

Another way to prevent inversion of the trench sidewalls is to increase the doping at the sidewall surface. Use of a direct, low-tilt-angle implant (e.g., boron, at an angle of 8° off the perpendicular to the top wafer surface, or at an angle of 82° off the perpendicular to the trench walls) is one reported way to do this (Fig. 9-9a).[16] However, since only one of the walls is doped by one implantation, the implant must be done four times, with the wafer rotated 90° after each step to ensure that all four walls of the trench are doped. In addition, this procedure is only effective for intra-well isolation (i.e., to isolate like-types of devices).

An alternative method for doping the sidewall of the *p*-well without doping the *n*-well with Si (thus making it suitable for CMOS inter-well isolation), was proposed by Davari *et al.*[17] In this approach, the *p*-well-sidewall-dopant is provided by boron diffusion from p^+-poly or borosilicate glass. The *n*-well does not require additional sidewall dopant, since phosphorus piles up during the brief thermal oxidation of the trench prior to CVD-oxide-refill.

In this selective-sidewall-doping process, a nitride/pad-oxide mask-stack is first formed and patterned. Next, after shallow trenches are etched by RIE to a depth of 0.3-0.5-μm, a boron diffusion source (*in-situ*-doped p^+-poly or borosilicate glass, BSG) is deposited. A lithography step is then used to etch the poly or BSG away from the *n*-well-sidewall and trench-bottom (see Fig. 9-9b). A thermal cycle is then carried out to drive the boron into the sidewall and bottom of the *p*-well trench. No boron is diffused into the active device areas since they are covered with nitride. The boron source is then etched away, and the remainder of the trench-refill process is conducted. The boron concentration at the sidewall-surface is controlled by the segregation during the brief thermal 50-nm-thick oxide-growth step. This approach was used for trench widths as small as 0.8-μm (i.e., used, for example in a 0.35-μm process). For the smaller-width trenches used in 0.25-μm and smaller technology nodes, the problem is handled

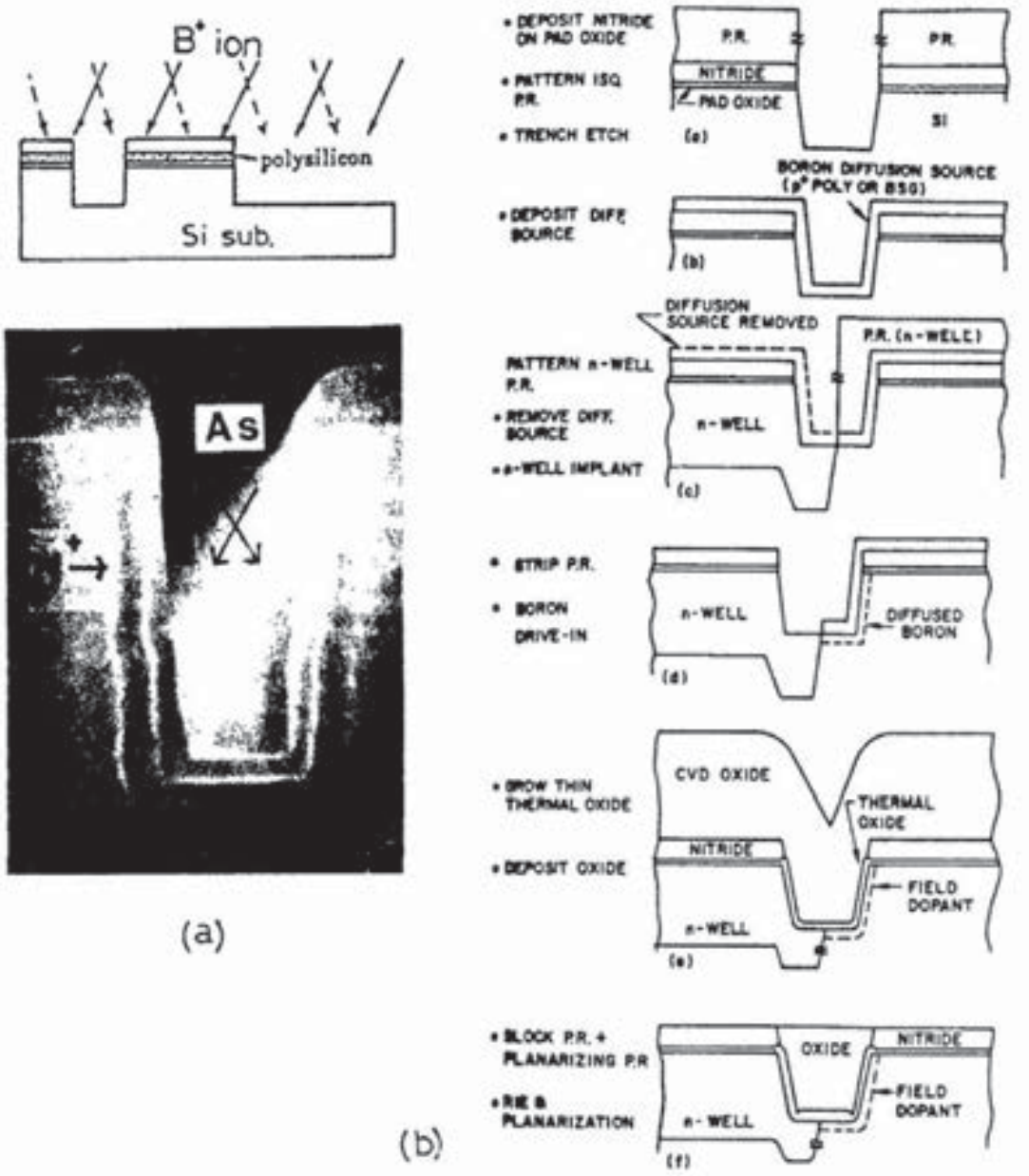


Fig. 9-9 a) Schematic of oblique ion implantation and SEM cross section result of this procedure.[16] (© 1987 IEEE) b) Shallow trench isolation process flow using a doped oxide as diffusion source of trench sidewall doping.[17] (© 1988 IEEE)

differently (e.g., the typical minimum trench width in a 0.18-μm process is about 0.4-μm with an active-area width of about 0.6-μm). That is, the p-well-doping implant-step (carried out after the STI structure is completed, using a high-energy-implant process) also places boron channel-stop-dopants at the bottom of the trench in sufficient concentration to prevent an inversion path from occurring

there. In addition, the boron-doping at the silicon surface in the channel-region of the NMOSFETs required to prevent DIBL in deep-submicron MOSFETs is high enough to keep the sidewalls of the trench from inverting.

9.4 DETAILS OF THE PROCESS FLOW TO FORM A BASELINE SHALLOW-TRENCH-ISOLATION (STI) STRUCTURE

In this section we present a process for fabricating STI structures. Note that the process sequence we describe is that of a baseline process, and one that is likely carried out for 0.25-μm CMOS generations. Some modifications can be made to this baseline process (even at the 0.25-μm technology node), and we will describe these variations in a later section. As ICs are scaled further, other modifications to the STI process will have to be made, and we will also discuss the need for such changes in the sections that follow this one that cover those process integration challenges in more detail.

The baseline sequence of steps for forming STI structures is given as follows:

1. The first several steps performed in fabricating an STI structure are the same as those used in the LOCOS isolation process sequence, with the exception being that the trenches for shallow trench isolation are the first structures formed on the bare silicon starting wafer. (On the other hand, in the conventional twin-well-CMOS process that used LOCOS, the wells were the first structures created on the bare-Si starting wafer.) In the STI process sequence, the bare, starting Si wafer is first cleaned (e.g., with an RCA clean), and then a pad oxide is thermally grown on the Si surface (typically to a thickness of 10-20-nm). Note that the role of the pad oxide layer is to prevent stresses that would exist if the silicon nitride layer (which is deposited next) were deposited directly on the bare-Si wafer. Such stresses could cause defects in the Si-substrate to occur. Note, however, that in the STI process the stresses associated with the formation of the STI structure are smaller than those involved in forming a LOCOS isolation structure. That is, in the LOCOS process, a long, high-temperature operation is performed to grow the field-oxide. During this field-oxide-growth step, the stresses between the nitride and the Si-substrate can be severe, and the pad oxide must therefore be sufficiently thick to keep these stresses from getting large enough to cause defects in the Si-substrate. In the STI process, there is no high-temperature field-oxide-growth process, and thus the stresses between the nitride film and the Si surface will not be as large as those in the LOCOS process. Hence,

the pad oxide in the STI process can be significantly thinner than the ones used in the LOCOS process (10-nm-thick vs. 25-40-nm-thick).

2. A silicon-nitride layer is then deposited on the pad-oxide layer (Fig. 9-10a). The silicon nitride plays two roles in the fabrication of the STI structure: **i**) it serves as a hard mask for etching the trench; and **ii**) it acts a CMP polish-stop-layer when the trench-fill dielectric material is planarized with a CMP process. The nitride-layer-deposition step is typically performed with an LPCVD process that deposits a silicon-nitride layer that is 100-200-nm thick. Note that the LPCVD of silicon nitride is a batch process that is simultaneously carried out carried out in a furnace tube on many wafers at once. As a result, the backsides of the wafers are also exposed, and silicon nitride is deposited on the back of the wafers (as well as on the front). Hence, this backside-nitride must somehow later be removed.

3. Then photoresist is applied and *Mask #1* is used to pattern the regions on the wafer surface where STI trenches will be opened.

4. The nitride and pad-oxide are then anisotropically etched, using a plasma-etching process (Fig. 9-10b). Usually this etch is carried out in a dry etcher designed for etching dielectric materials (*Etcher #1*).

5. The trenches are then etched (Fig. 9-10c). This silicon-etch step is also carried using a plasma-etching process. The goals of this etch process include: **a**) etching the trenches to the correct depth; **b**) controllably producing trenches with the sidewalls having the desired slope (e.g., from 70-85°, depending on the technology node, the trench depth, and the type of circuit being fabricated); **c**) producing trenches with rounded bottom corners; and **d**) having the etch process produce smooth sidewalls and minimum Si damage.

In general, it is easier to fill trenches without forming voids in the trench-fill material if the trench openings are sloped. Obviously, this requires more Si area for the trenches, and thus as ICs are scaled, the sidewall profiles will have to become more vertical. Gasses used for etching the trenches are mixtures of HBr and other gases (e.g., $HBr/Cl_2/O_2$, or $HBr/NF_3/He/O_2$).[18]

The trench-etch process is usually carried out in a second etcher (*Etcher #2*), different from the one used to etch the nitride and oxide layers (*Etcher #1*). That is, *Etcher #2* is usually a modified version of a gate-etch tool (i.e., one in which polysilicon is etched).

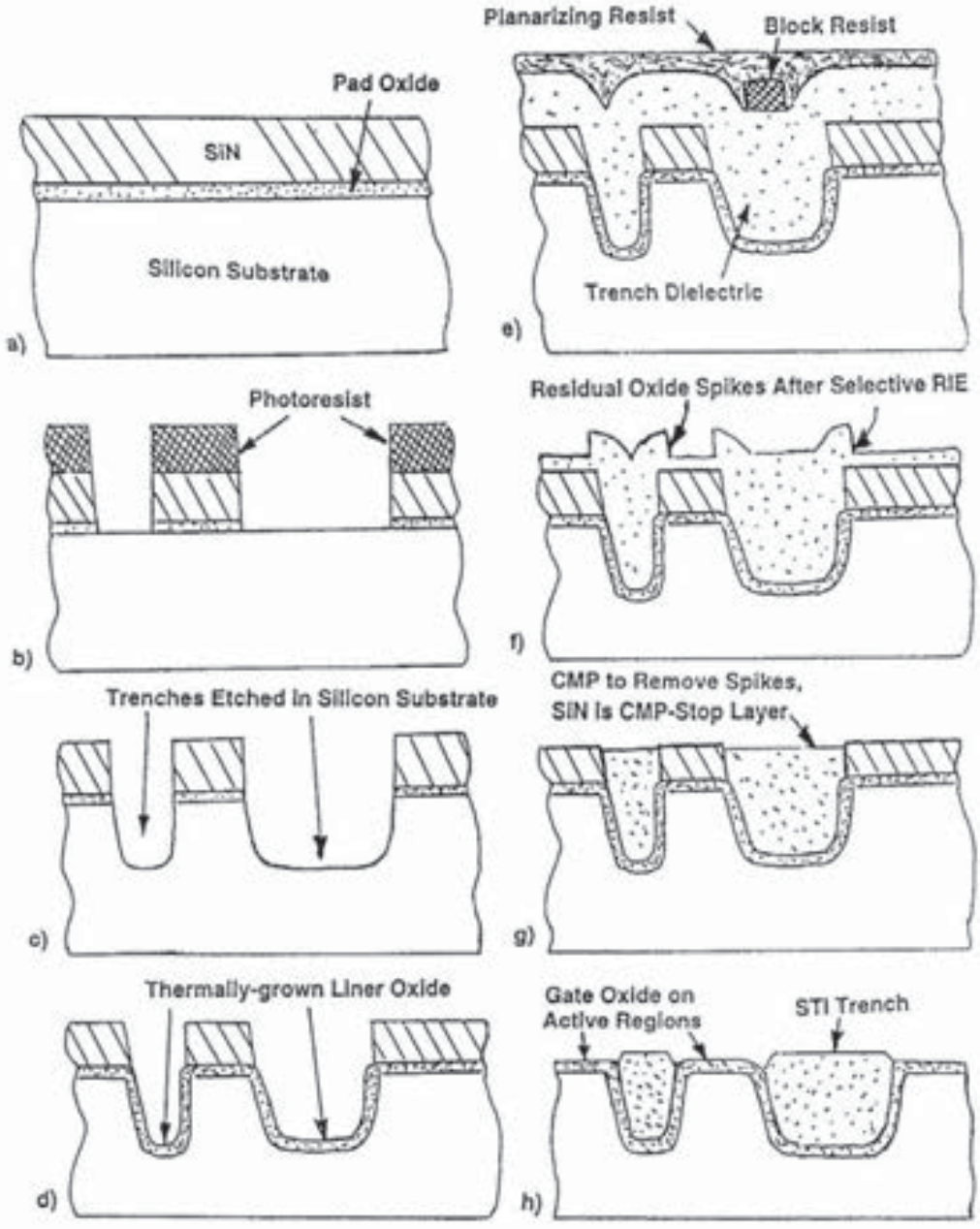


Fig. 9-10 Baseline process flow used to form an STI structure: a) Form pad oxide by thermal oxidation, then deposit a silicon-nitride layer by LPCVD; b) Pattern the nitride and pad-oxide layers; c) Etch trenches, strip resist; d) Grow liner-oxide on the trench walls, using thermal oxidation; e) Deposit the trench dielectric & densify it, then deposit a block resist layer & pattern it. Finally deposit the planarizing resist layer; f) Use a selective-RIE-etch process to simultaneously remove the resist and the trench-oxide-layer, leaving residual oxide spikes; g) CMP the CVD-oxide-layer to abrade spikes and planarize its surface. The silicon nitride acts as a CMP-stop-layer; h) Strip the nitride and then grow the gate-oxide on active regions. The STI structure is now completed.

Note, however, that these two etch steps could also be carried out in a single etcher. Significant enhancements to process productivity (in the form of higher throughput and lower cost) could be obtained if the opening of the silicon-nitride hard-mask-layer and the etching of the trenches were carried out *in-situ* in a single etcher (and such a process has been termed *in-situ hard-mask-open*). Etch tools capable of *in-situ* hard-mask-open are available from Applied Materials, Lam Research, and TEL America, and some such processes have reportedly been put into production by 2000.[19] The main challenge of successfully implementing an *in-situ* hard-mask-open process is selectivity to the photoresist.[20]

The trench etching process is difficult for a number of the following reasons:

a. The trench depths across the wafer must have uniform depth to ensure correct circuit operation (the dependence of circuit performance on trench depth will be discussed later, in the section that deals with CMP-dishing problems). However, some dry-etch processes exhibit *aspect-ratio-dependent-etching* (ARDE) characteristics, in which material from wider-spaced trenches is etched more rapidly than from narrower trenches. Thus, etch-process development must be employed to find etch-process conditions in which ARDE effects are limited.

b. The trench etch process is a 'blind' process, in that there is no etch-stop layer at the bottom of the trench, nor is there any material present at the trench bottom that will create an end-point-signal. Hence, an etch process must be available that is so well-controlled



Fig. 9-11 SEM of etched trenches of an STI structure with trench walls having an 87° slope and bottom corner rounding. Courtesy of Lam Research.

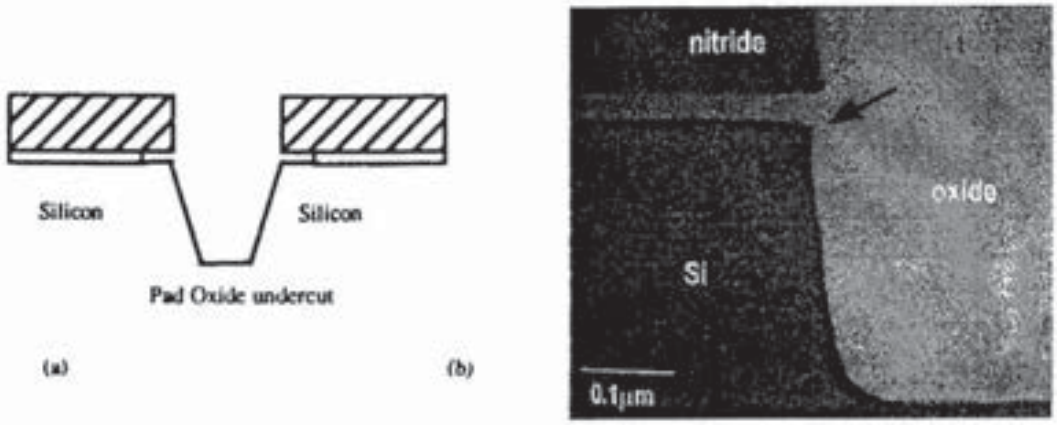


Fig. 9-12 a) Pad-oxide trench-undercut in the STI-baseline process. b) TEM of an as-etched trench profile showing the sharp corner of the active Si region at the top of the trench.

that a timed etch can be relied upon to produce trenches with the correct depth, from wafer-to-wafer, and across a wafer.

c. The etch process should yield smooth trench sidewalls with angles of 70–85°, rounded bottom-corners (to minimize stress), and a residue-free silicon-surface after etch (see Fig. 9-11). Trenches with rounded-bottom corners minimizes stress to the silicon lattice encountered during the trench-liner-oxidation process of Step **8** and the trench-filling-process of Step **9**. It was reported that using gas etch mixtures consisting of $HBr/NF_3/He/O_2$ gave improved corner rounding results than mixtures containing HBr/Cl_2.[21]

6. After the trenches are etched, the resist is stripped.

7. The pad oxide is next undercut with a short HF dip (Fig. 9-12a). The reason this step is carried out has to do with the fact that upon its completion, a silicon dioxide layer is thermally grown on the exposed trench surfaces. This so-called *liner-oxide* layer serves several important roles. First, the process of growing this oxide consumes any residual damage caused by the plasma-trench-etch process. Second, the thermal oxide passivates the Si surface of the trench, providing as stable interface between the Si and the trench fill material. (The threshold voltage of the parasitic field MOSFETs must be kept high enough so that leakage paths at the Si-surfaces of the trenches are not electrically activated during normal device operation. To ensure that this requirement is met, the trench-Si/filler-material interface must be electrically stable. A thermal oxide in

contact with the Si surface provides the most electrically stable of interfaces.) Third, at the top corner of the active-region (i.e., where the active-region and the isolation-region meet, see Fig. 9-12b) the Si normally has a sharp corner after the trench etching has been completed. Such a sharp corner will enhance the electric field lines at this corner and degrade the MOSFET turn-off characteristics (as will be discussed in more detail later). To suppress this effect, the top-sharp-corner of the silicon active-region must therefore be rounded. Device simulations have demonstrated that the parasitic double-hump in the I_D-V_{GS} curves can be eliminated by having a trench corner with a radius of 50-nm if a trench with no more than an 80° sidewall slope is specified.[22] This can be done in a number of ways, but here, we will assume that such a "corner-rounding" process is also performed by the thermal-oxidation step that forms the liner-oxide.

It has been demonstrated that the top-sharp-corner will be more effectively rounded by the thermal-oxidation step that forms the liner if the pad-oxide is undercut prior to performing the thermal-oxidation liner-step. Hence, as shown in Fig. 9-12a, this pad-oxide-undercutting step is performed next, i.e., by dipping the wafer in HF to etch-back the pad-oxide to a controlled extent.

8. The thermal-oxidation step that forms the liner oxide and rounds the sharp-top-corner of the trench is performed next. The oxide grown in this procedure is typically 40-50-nm thick, and it is grown in dry-O_2. It has also been reported that maximum corner-rounding is achieved if the liner-oxide is grown with a high-temperature (1050-1100°C) rapid-thermal-oxidation (RTO) process (Fig. 9-13).

Note that a thin layer of the nitride is converted to oxide during this liner



(a) Lower temp+Hcl (b) High temp RTO

Fig. 9-13 Cross-sectional TEM of the top trench corner after pad oxide undercut & liner oxidation: HCl in the ambient (a); and higher-temp RTP-oxidation (b) improve corner rounding.

oxidation step, an effect that must be taken into account when the nitride is later stripped. Also note that in DRAMs fabricated with STI, a thin silicon-nitride liner may also be deposited by CVD prior to filling the trench with a CVD oxide. This nitride liner helps decrease leakage current, a critical parameter in DRAM operation.

9. Next, a CVD-dielectric film is used to fill the trenches. Since the deposition is a blanket process, some of the material is deposited onto the active regions that are still covered with nitride (i.e., 8000-9000-Å of gap-fill dielectric are deposited when filling a 5000-Å-deep [0.5-μm] trench). As a result, the material must be removed from the active regions after the deposition process is completed (Fig. 9-10d). As mentioned earlier, a number of CVD processes have been pursued for filling the trenches without void formation, including: LPCVD TEOS; (sub)atmospheric-pressure CVD of ozone/TEOS; and high-density plasma CVD (HDP-CVD).

10. The dielectric material is densified at about 900°C. This step is carried out so that the trench-fill dielectric will exhibit etch rates comparable to that of thermal SiO_2. This characteristic is necessary so that excessive loss of the trench-fill dielectric is not incurred during later HF deglazing steps (i.e., steps **15**, **17**, and **20**). This densification process reduces the wet-etch rate by about 20%.

11. A photoresist layer is next spun-onto the wafer. This resist is then patterned to form a reverse-mask of the active regions and a planarizing-resist layer is deposited over the patterned reverse-masked wafer (Fig. 9-10e). The oxide and the resist are etched simultaneously in an RIE-etch-back step that etches resist and oxide at the same rate. This reduces oxide thickness in the field areas. The result of partially planarizing the surface with this REI step is that the time needed to planarize the wafer surface during the subsequent CMP step is decreased. A shorter-duration CMP step decreases the problem of dishing (as discussed further in Sec. 9.5.3).

12. After the resist is patterned and the planarizing-resist layer is deposited, the wafer is etched back with an RIE process described in Step **11**, and as shown in Fig. 9-10f.

13. Any remaining photoresist is then stripped. Small spikes of oxide, however, remain near the edges of the active/field regions (see also Fig. 9-25). These must be removed to produce a globally planarized STI surface.

14. A CMP-step is used to polish back the dielectric layer to remove these

spikes, which is continued until the nitride is reached (the nitride acts as a CMP-stop layer, Fig. 9-10g). A major issue involved with this CMP-step is dishing of the trench-fill-dielectric in wide trench regions. The reverse-mask procedure of Steps **11,12,** and **13** is one method used to mitigate the dishing problem. Other dishing-suppression techniques are discussed in Section 9.5.4.

15. The nitride is then stripped, using a hot-phosphoric acid, wet-cleaning process. But, it should be noted that the stripping process is actually done using two wet-etching steps. That is, the thin layer of nitride that was converted to an oxynitride layer during the trench-liner-formation thermal-oxidation-step must be removed prior to carrying out the nitride-stripping-step itself. The oxidized layer on the nitride will not etch in hot phosphoric acid, and unless it is removed, the phosphoric acid will thus not strip the nitride. Thus, a short HF-etch step is first performed to remove the oxide layer on the nitride. Note, also that the CVD-oxide that has filled the trench is also attacked by this HF-etch-step. Since the CVD-trench-oxide etches more rapidly than does the oxide layer on the nitride, this HF-etch-step must be carefully optimized. Another factor that should be considered when determining this HF-etch is the possibility of there being insufficient CVD-oxide removal from the nitride by the CMP process. If the oxide on the nitride remaining after the CMP step is too thick, the HF-etch of the process in this step may not remove all of it, making the nitride-layer-etching process ineffective, leading to potentially significant yield degradation.

16. After removing the oxide layer on the nitride, the nitride can be stripped by the hot phosphoric acid. There are several reasons why a wet process is used to strip the nitride, rather than a dry process. First, the wet process is inexpensive and selective (i.e., the hot phosphoric acid will not attack the underlying pad oxide, or Si substrate). Second, a wet process will remove the nitride deposited on the back of the wafer earlier by LPCVD, as well as the nitride on the front wafer surface.

17. Normally the pad oxide is stripped in HF at this point. Again, the oxide in the trench is also attacked by this HF etch step. This etching plays an important role in forming the topography at the edge of the active region, where the polysilicon gate is deposited on both the active area and the trench region (see Fig. 9-14). This topic will be considered again later.

18. Grow sacrificial oxide, about 20 nm thick, to protect the exposed Si surface of the active areas from contamination during the implant process.

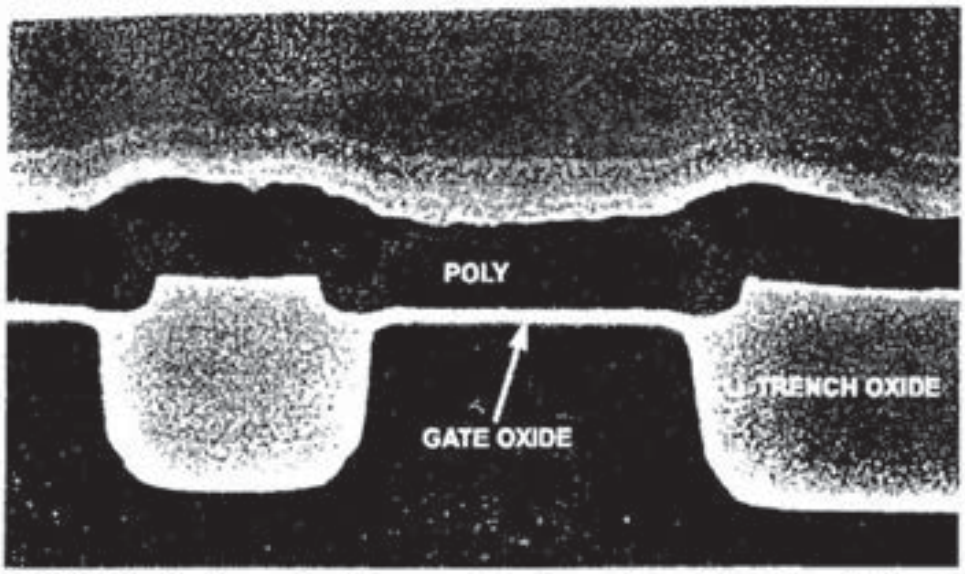


Fig. 9-14 SEM showing polysilicon crossing the region at the edge of the active device region and the trench in an STI structure.

Note that another reason for stripping the pad oxide and growing a sacrificial oxide is that an effect similar to the Kooi-effect (that occurs during LOCOS isolation - see Vols. 2 and 3 for more details on the Kooi effect) has also been reported to possibly occur during STI processes. That is, if the trench-oxide is densified in an ambient containing O_2, this O_2 can react with residual H_2 in the trench oxide film or in the SiN layers present on the chip to form H_2O. This H_2O in turn can react with the SiN to form NH_3 which would lead to nitridation of the Si as in the Kooi effect. The net result of such a phenomenon would be that the gate-oxide grown where such a nitrided region existed could be thinner there, as shown in Fig. 9-15.[23] To avoid the possibility of such gate oxide thinning (and a resultant degradation of the gate oxide integrity [GOI]), a sacrificial oxide is grown which will consume any

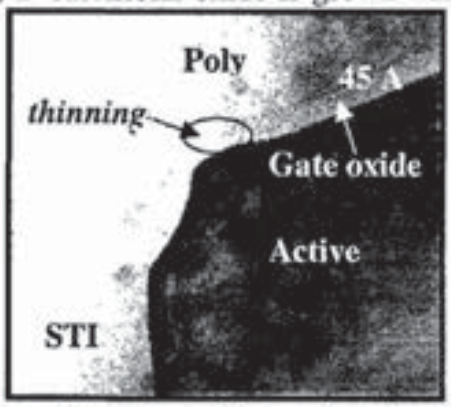


Fig. 9-15 TEM showing thinning of the gate oxide near the edge of an active region in an STI structure due to the Kooi effect.[23]

nitrided regions that formed on the active Si surface regions.

19. The wells are now created by performing a high-energy-implant to place dopants within the Si substrate in the active regions, and beneath the STI structures to produce a channel-stop layer of dopants. In addition, the V_T-adjust- and punchthrough-implants are carried out at this time.

20. The sacrificial oxide is then etched, again using a wet (HF-etch) process. Once again, the oxide in the trench is subjected to an HF etchant, further reducing the thickness of the oxide in the trench (and also further modifying the trench edge corner region - Fig. 9-16). By the way, it should not necessarily be construed that the HF etches will remove enough oxide in the trench regions to cause the trench oxide to be recessed. While this might be a possibility, not much oxide thickness loss occurs as a result of the HF etches. The more significant effect is on the region of the trench oxide near the trench top corner.

21. Finally, the gate-oxide growth process is carried out (Fig. 9-10h).

9.5 ISSUES INVOLVED WITH STI PROCESS INTEGRATION

9.5.1 Issues Involved With Etching Trenches for STI

Some of the issues involving trench etching for a baseline-STI structure were discussed in Section 9.4. Here, we elaborate on some other issues that arise as more advanced STI structures are needed. In the baseline-STI structures, the trench sidewalls were typically sloped, with slopes ranging from 70-85°. As the trench widths are scaled to ever decreasing dimensions, the trench sidewall profile requirements are becoming steeper, with slopes as high as 88° being called for in some applications.[24]

One consequence of this change is that the *in-situ* hard-mask-open etch process may become more feasible. That is, in the baseline process of Sect. 9.4,

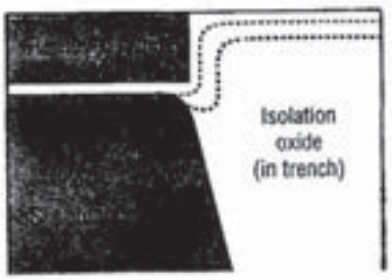


Fig. 9-16 Topography of the top corner of the Si active region and the oxide surface in the trench. The dotted lines represent the removal of oxide that occurs during the wet HF dips before the sacrificial oxide and gate oxide growth steps.

STI structures had sidewall profiles with smaller slopes (e.g., 75-80°). If the trenches were etched while the resist remained on the wafer, the polymer deposition on the sidewalls during the trench etch would lead to increased ARDE effects, and this would have contributed to trench-depth and profile-variations between trenches of different widths. Thus, the two-etcher trench-etch-processes (with the resist being stripped after the 1st trench-etch step) were more desirable. To produce steeper trench profiles, lower-pressure etch-processes with higher ion-bombardment-fluxes are used. In these high-ion-bombardment-flux processes, there is much less polymer deposition needed on the trench sidewalls to protect them from lateral etching. As a result, the presence of the resist during the trench-etching stage can be more tolerated, opening the door for the *in-situ* hard-mask-open single-etcher solution. Note that the steep trench-sidewall profiles are also controlled by using low-pressure, high-density-plasma etchers. The gases employed seek to exploit a non-depositing etch process, with gas mixtures for example consisting of Cl_2 and Ar.

Another trend as devices are scaled, is toward using shallower trench depths. The baseline-STI structures used trenches with depths of ~0.5 μm. As devices approach the 0.1-μm node, the trench depths are expected to be decreased to < 0.2-μm, at least for some circuit applications at that technology node.[25] The trench depth is set by two main considerations:

1. The oxide thickness in the trench (field-oxide), together with the channel-stop-doping under the field-oxide determine the threshold voltage of the parasitic-isolation-region-MOSFETs. To keep this threshold-voltage high enough to prevent significant isolation leakage, a minimum field-oxide - thickness (i.e., trench depth) is needed. As the operating voltages are scaled for future device generations, smaller threshold voltages of the parasitic-isolation-region-MOSFET will be acceptable. Thus, a shallower STI-trench will be feasible.

2. The trench-oxide-depth impacts the capacitance value of the polysilicon and 1st-level-interconnect-metal to the substrate. The thinner the trench depth, the higher the capacitance. However, use of smaller operating voltages will also allow thinner STI-trench-depths to be used, with the exact depth value to be determined by the circuit application. Note that the severity of dishing problem, to be discussed in Sect. 9.5.4, will also play a role in determining how much the trench depth will be permitted to be reduced.

The contamination caused by the trench-etch-process is also an issue. Since an isolation structure is being fabricated, concerns regarding metallic contamination

increase with ever-decreasing device dimensions. To address this problem, Lam Research reportedly replaced all ceramic chamber components with quartz (as the ceramic material may contain trace levels of metal that could be released into the etch-gas ambient during the etch process, and then these could redeposit on the exposed Si of the trench walls). Applied Materials attacked this question by implementing a Clean Chemistry™ using fluorine throughout the etch, which eliminates the need for frequent dry-cleans of the chamber.

As already mentioned in the baseline-STI process sequence, maintaining *trench-depth-uniformity* across a die (and from wafer-to-wafer), is a challenging problem. However, merely controlling the *trench-depth* is also particularly important, because the uniformity of fill and also uniformity of CMP, in turn, depend on trench-depth-uniformity. The trench etch is essentially a blind etch, with no etch-stop-layer, nor any material differences in the Si substrate regions available to provide an end-point-detection signal. Thus, in the 2000-timeframe, the etch-depth-control must rely only on a timed-etch process.

Depth- and profile-control-nonuniformity are functions of reactor design and etch-gas chemistry. The stability of the chamber conditions also affect the trench-etch-rate-uniformity. The etch-rate will vary for different patterns, structures, open areas, and different microloading conditions. Thus, the specific etch-rate may need to be re-calibrated, depending on a particular set of conditions being faced. This indicates that there is currently a great opportunity for a production-worthy endpoint-detection-system for the STI-trench-etching process.[26]

9.5.2 Corner-Rounding Techniques in STI Structures

In Steps **7** and **8** of the baseline-STI process we described how the top of corner of the STI trench must be rounded so that adequate leakage performance charac-

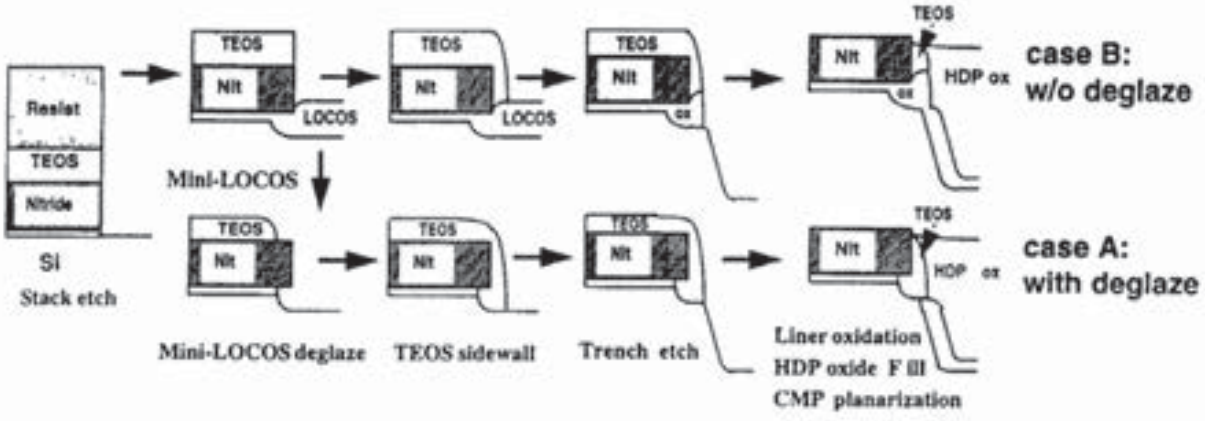


Fig. 9-17 Process sequence of STI using mini-LOCOS edge.[29] (© IEEE 1997)

teristics will be exhibited by CMOS ICs that employ STI. The corner-rounding process described in the baseline-STI flow involved undercutting the pad-oxide just before performing the rounding (aka *liner*) oxide procedure. The best corner-rounding results are obtained if an RTO step in dry-O_2 at 1050-1100°C is used.[27] In the paper that described this step, several other corner-rounding procedures were also studied. These included: **a**) adding HCl to the dry-O_2-RTO process;[28] **b**) carrying out a post-CMP-oxidation in dry-O_2 at 1050-1100°C;[29] and **c**) using a "mini-LOCOS" process (see Fig. 9-17).[30] Good rounding is observed in the post-CMP-oxidation-process, but it comes at the expense of reducing the active-device area to some degree. The mini-LOCOS process ads considerable process complexity. A hydrogen-annealing procedure was also reported as an alternative method of achieving corner rounding. In this approach, the wafer is annealed at 1100°C after the trenches are etched (but before being filled). It is proposed that at this temperature, Si-surface atoms migrate and establish positions that represent the structure with the lowest free energy (namely a sharp corner becomes a rounded corner, see Fig. 9-18).[31,32] A novel approach to rounding the trench corner during the trench process itself is described in Sect. 9.5.6.

9.5.3 Trench-Fill Dielectrics for STI

We know that the trenches of the STI structure are filled with a dielectric, and that deposition of oxide-like films by CVD has been the chief technique for

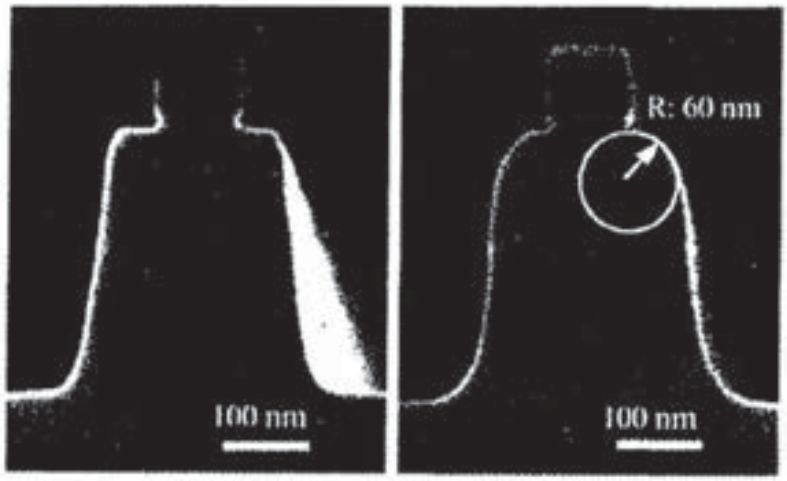


Fig. 9-18 Cross-sectional SEM of before (a) and after (b) H_2-annealing. Annealing condition was 950°C, 380 torr, 60 sec. Silicon migration phenomenon can round out STI's top-corner-rounding-radius up to 60-nm.[32] (© IEEE 1998)

filling the trenches. A report published by Nag *et al.*,[33] compared several such dielectrics deposited by such methods for their suitability as the trench insulating material. The criteria they used to evaluate these materials were as follows:

1. The deposited material must be void-free, to avoid opening seams during HF-deglazing steps (i.e., such voids could later lead to poly filaments).

2. Post-deposition HF-deglazing also reduces the trench-fill oxide thickness. This can expose silicon at trench corners and cause leakage. Therefore a comparison of the etch rates of the trench-fill-oxides relative to thermal oxide must be established.

3. Film stress, shrinkage, and composition after the high-temperature anneal (densification) affect the integration.

4. The CMP-polish-rate relative to nitride and polish uniformity of the oxide impacts the trench-fill-oxide thickness remaining after CMP that influences the "corner effect" problem.

5. Isolation characteristics of PMOS and NMOS diodes have to be tested to verify the electrical performance of trench-fill dielectrics.

In this study four films types were compared: LPCVD TEOS; HDP-CVD oxide; SACVD O_3-TEOS; and an HSQ-SOG (see Vol. 1, 2nd Ed for more information on HSQ-SOGs, as well as Chap. 14 of this text). In general, they concluded that the HDP-CVD oxides were the most suitable of these four films for STI-trench filling. Table 9-1 summarizes the results of their study. Trenches 0.5-μm-deep and widths as small as 0.22-μm could be filled without forming voids.

Table 9-1 TRENCH-FILL DIELECTRIC EVALUATION CRITERIA

Criteria	SACVD	LPTEOS	HDP	SOG
Gap-fill	OK	No	OK	OK
Hi-T Anneal required	Yes	Yes	No	Yes
As-dep'd HF-etch rate vs. thermal ox.	5	2.1	1.1	
Densified (850C) HF etch	1.3	1.1	1.1	1.1
CMP rate vs. SiN	5.1:1	5.1:1	5.1:1	5.2:1
Shrinkage at 850C	5-6%	3-4%	0-0.8%	22-28%
Moisture Absorption	Yes (cure)	No	No	Yes (cure)
Process	Seed/Cap	1 step	1 step	Seed/Cap

Some more recent reports, however, have indicated that some of the other films might still be good candidates as trench-fill materials. For example, a study by Xia *et al.*, demonstrated that a two-step deposition of a BPSG film using a sub-atmospheric (SA)CVD-process of O_3-TEOS is also capable of void-free gap filling of trenches whose widths are as small as 0.06-μm and whose depths are 0.35-μm deep.[34,35] Figure 9-19 shows a SACVD chamber used to deposit such films.

Nevertheless, since the HDP-CVD method appears to have emerged as the technology of choice for depositing STI trench oxides, some comments should be made about this process (and the potential impact that the HDP-CVD method can have on the STI structure). Note that Novellus (SPEED® HDP-CVD model) and Applied Materials are vendors of HDP-CVD tools for dielectric deposition.

The basis for the improved gap-filling capability by films deposited with HDP-CVD is that HDP sources provide a high-density of low-energy ions to the wafer surface. These ions simultaneously sputter the film being deposited by CVD. The sputtering removes film material from some regions on the surface more rapidly than from other regions. This effect is exploited to yield films that fill narrow gaps without voids. Another benefit of HDP-CVD is that films can be deposited at lower temperatures. This mechanism is described in detail here.

The film being deposited by CVD initially covers all the surfaces of the wafer conformally (including the sidewalls and bottom of any trenches or contact holes). As the film grows thicker, an overhang develops at the top of the trench. This will close-off the opening of the trench before the dielectric material completely fills it, leaving a void. If the overhang can be prevented from

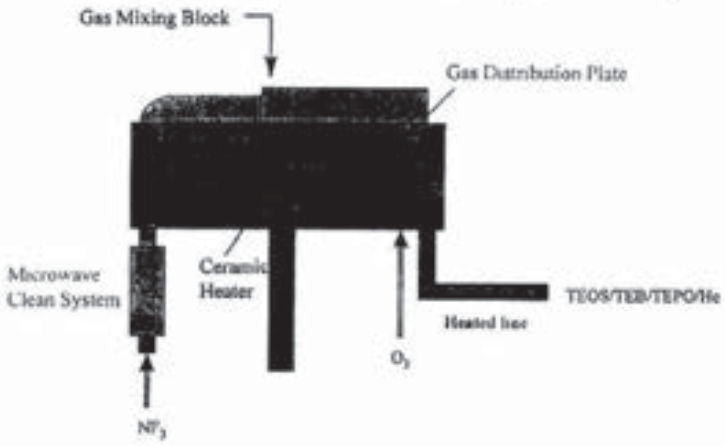


Fig. 9-19 Schematic of Applied Materials SACVD Giga-Fill BPSG chamber, equipped with a high-temperature heater and remote microwave system for predeposition cleaning.[34]

forming, it may be possible to fill the trench without void formation. This is accomplished in an HDP source by the bombardment of the wafer surface with a high density of ions. For depositing oxides by HDP-CVD, a mixture of SiH_4, O_2, and Ar (or He) gases is most commonly used. The HDP source excites the oxygen and argon gas molecules. A bias applied to the wafer pulls the energetic ions out of the plasma and directs them toward the wafer surface. The oxygen reacts with silane to form SiO_2 while the Ar ions simultaneously sputter deposited material away. A high-density-plasma source is needed to maintain a high-ion-density at the relatively low pressures (a few mTorr) needed to obtain the directionality required for good gap-fill. The good gap-fill is achieved as described below.

The sputter yield is higher for surfaces that face the incoming ion beams at an angle, than at those surfaces perpendicular to the incoming ion directions. The maximum sputter-etching-rate occurs on surfaces at about 45° to the normal. Thus, a deposition-to-etch-ratio exists that varies with location on the wafer surface. The net deposition-rate is higher on flat surfaces (the field-regions of the wafer and the trench bottoms), than it is on the trench corners. (These angular surface regions are the 45° facets formed by depositing dielectric film at the top of the trenches.) By adjusting the deposition-to-etch-ratio, the trench openings can be kept from closing up while the dielectric film fills the trenches from the bottom up (Fig. 9-20). Sputtering from the via bottoms also causes re-deposition of some material onto the sidewalls. This helps re-distribute the film so the gap is

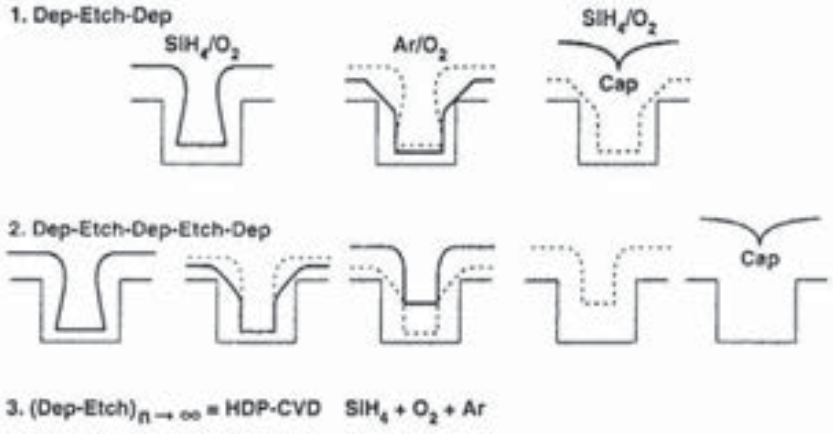


Fig. 9-20 High-density plasma CVD achieves good gap filling through simultaneous deposition and etching. The deposition-to-etch-ratio is usually kept between 0.14 and 0.33.[43] Reprinted with permission of Semiconductor International.

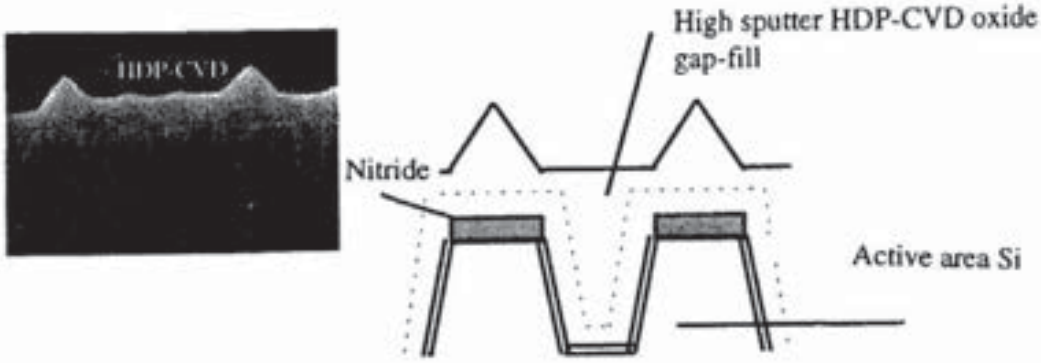


Fig. 9-21 The surface topography of a dielectric film deposited by HDP-CVD showing the sharp spikes that remain as a result of the dep-etch nature of the HDP-CVD process.

filled without forming a void. To achieve good filling, the ratio of deposition to removal rate should be no larger than 3.0. The wafers are also heated by the ion bombardment. Thus, the deposition-rate is limited by the highest sputter-etch-rate that can be tolerated. The process is run at low pressures (< 10 mtorr), which allows ions striking the wafer surface to cross the sheath with no collisions. They arrive with the full energy they receive from the voltage drop across the sheath.

At the conclusion of the deposition process, the surface of the dielectric will generally still contain sharp ridges over the underlying metal lines (Fig. 9-21). CMP is needed to restore the wafer surface to a planar condition (although the amount of CMP required is reduced from that needed for a dielectric film deposited with a non-HDP-CVD process).

One issue that arises with HDP-CVD is that the combination of deposition and etching can sputter the top of the nitride corners, in effect "clipping them." One technique that suppresses this effect is to use He instead of Ar as the sputtering species in the HDP-CVD process. Another approach is to decrease the deposition-to-(sputter)-etching ratio of the process. Since this reduces the gap-filling capability, a trade-off must be made to optimize the overall deposition process.

9.5.4 Dishing Problem Associated with the CMP of STI Trench Dielectrics

Another CMP problem that occurs when fabricating STI structures is *dishing*. The effect involves the thinning of trench oxide regions that are exposed during the overpolish time that occurs after the CMP-stop layer is reached (Fig. 9-22). A silicon nitride layer is used to act as a CMP-stop layer. The thinning of the exposed oxide areas between the nitride-CMP-stop features is known to be highly dependent on feature size, with the dishing effect becoming larger as the

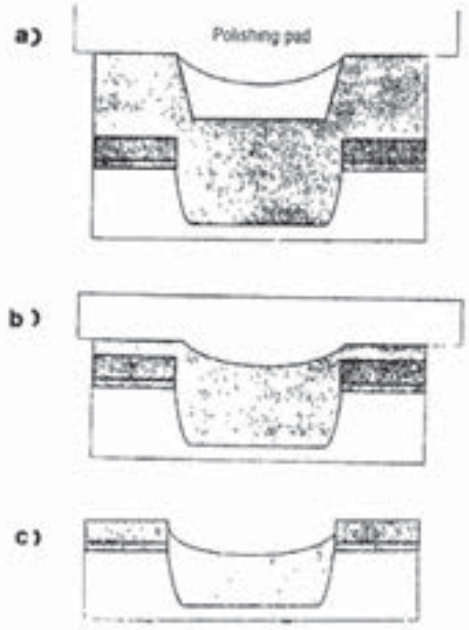


Fig. 9-23 Origin of the dishing effect in CMP: (a) polishing pad flexes into the trench opening; (b) formation of a concave surface; (c) trench area remains "dished" after planarization. Reprinted with permission of Solid State Technology. Published by PennWell.

width of the field oxide is increased (Fig. 9-24). For example, in the CMP process of Fig. 9-24, dishing causes the field oxide thickness to decrease by more than 50-nm when the field oxide width exceeds 100-μm.[36] Dishing occurs because the polishing pad is not perfectly rigid, and because oxide removal is faster than nitride removal. If a pad exhibited perfect rigidity, CMP of the exposed oxide regions would produce a perfectly planar surface (because the exposed recessed regions would not be subjected to any polishing pressure). However, since actual pads exhibit some flexibility, above-zero polishing pressures are produced in recessed areas (especially in the middle of wide exposed regions) when polishing continues after the nitride has been exposed. Thus, some oxide is removed from these regions, and a "dished" region is observed after the polish step is completed. To minimize dishing, the silicon dioxide polish should be stopped as soon as the oxide over the nitride clears. Accordingly, the quality of an STI-CMP-process is measured by the control of trench-oxide and nitride thickness. This requires tight control of the step height between the trench-oxide surface and silicon surface after nitride strip.

Dishing must be minimized to maintain planarity for subsequent wafer processing. It still represents a significant limitation of the STI process. A number of

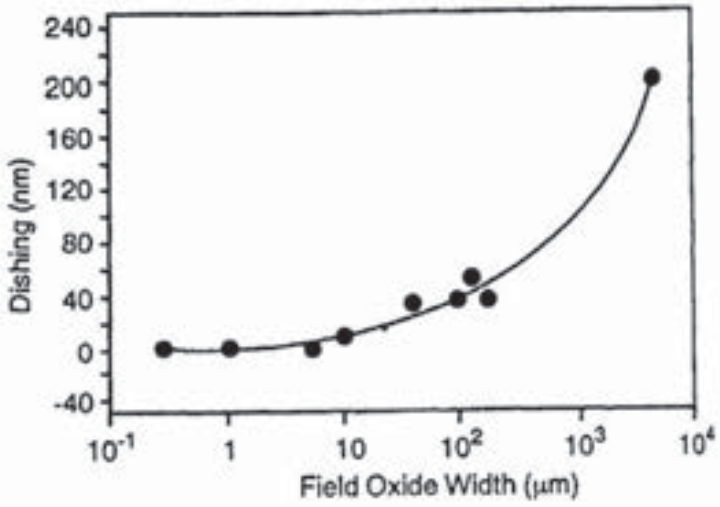


Fig. 9-24 Dishing versus field-oxide-region width for a specific oxide CMP process[36] (©1993 IEEE).

approaches have been proposed to decrease dishing, including: **1**) a combination RIE/CMP process (as described in Sect 9.4); **2**) dummy active-regions; **3**) adding an extra nitride layer onto the trench-dielectric after it has been deposited; [37,38] **4**) adding an extra nitride layer onto the dielectric after it has been deposited, and using a reverse block-mask;[39] **5**) implementing a direct-polish CMP-process using either a linear-polishing tool, or a rotating-polishing tool with a harder polishing pad and an efficient endpoint-detection-system.

As described earlier, in the *RIE/CMP approach* the goal is to reduce the oxide thickness over active areas prior to CMP, so that a shorter CMP-oxide-removal step is needed. Since CMP is performed for a shorter time, the degree of dishing in wide trench regions is reduced. In the first step of the RIE/CMP process, a mask that contains a pattern that is the reverse of the trench-opening mask is used (but, with the exception that active-regions below a specified nar-row width are also not exposed - Fig. 9-24).[40] This mask is aligned so that its opaque patterns cover the trench regions on the chip surface. That is, this mask exposes the active regions of the chip, leaving the trench regions covered with resist (Fig. 9-10e). After the subsequent deposition of another planarizing resist layer (see Fig. 9-24), an etchback step is carried out using RIE. This process re-moves resist and dielectric material at the same rate, and so thins the oxide covering the active regions. Note that oxide in the trenches is not etched while oxide over the active regions is being removed, because the trench oxide is still covered with resist.

After this RIE step the resist is stripped, leaving the oxide over the active areas partially removed, Now, when the CMP step is performed, it only has to remove a thinner oxide layer, reducing the CMP time This not only reduces

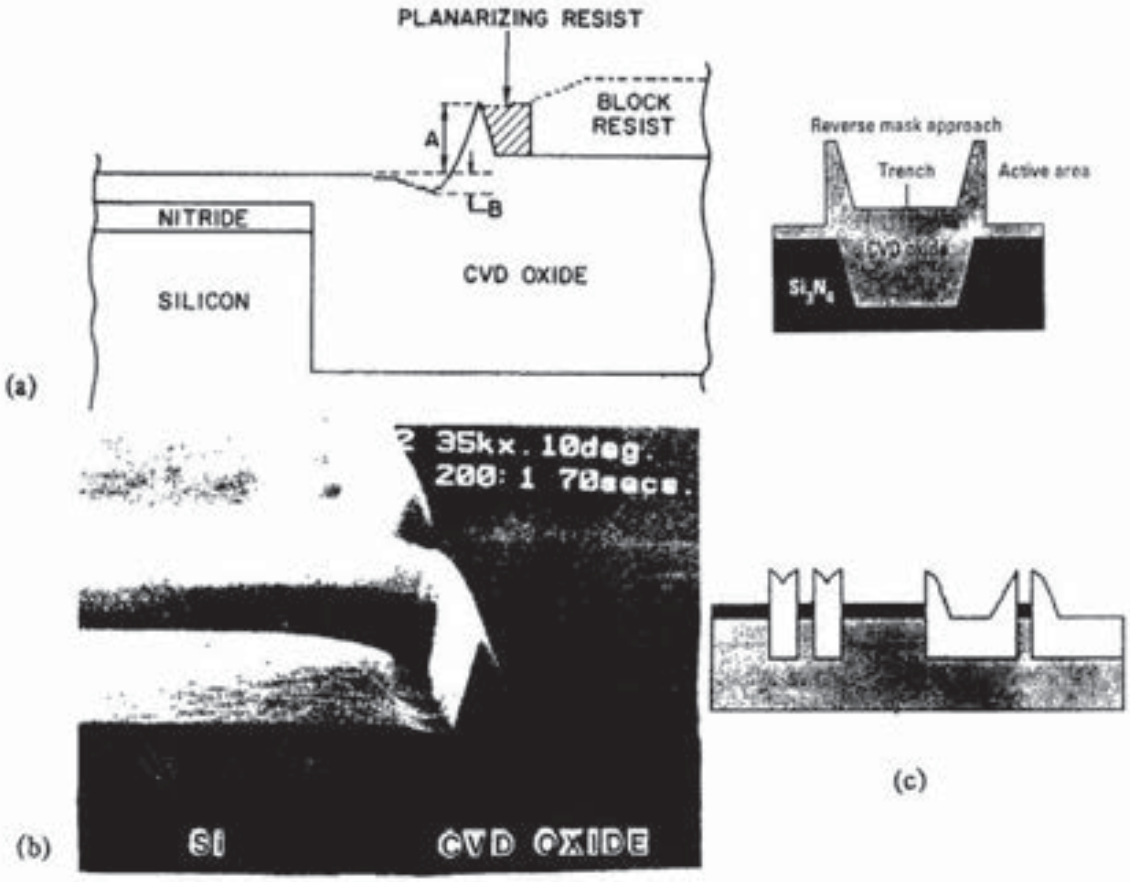


Fig. 9-25 Spikes in the dielectric layer after the RIE step of the RIE/CMP process: a) Simulated cross-section after RIE; b) SEM cross-section after RIE and resist strip, showing dielectric spikes; c) Schematics of the dielectric spikes after RIE and resist strip.

dishing, but reduces CMP cost (since a shortened CMP step allows the throughput of the CMP procedure to be increased). Note also that dimensions on the reverse-trench mask are actually slightly smaller than the dimensions of the active-regions themselves. Thus, even when the mask is aligned to the active-regions, a small border of the active-region oxide remains covered by the resist. As a result, at the conclusion of the RIE step, small "spikes" of oxide remain on the wafer surface (Fig. 9-25).[41,42] But, these spikes can be effectively removed by the CMP step, allowing a planarized oxide surface over the chip to be obtained. The main drawback of the RIE/CMP method is that it increases the complexity (and thus also the cost) of the process used to planarize the trench oxide.

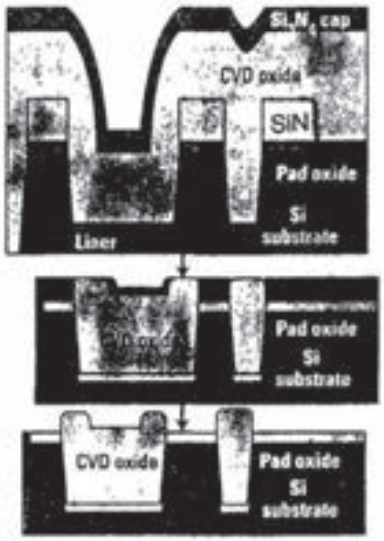
In the *dummy-active-region approach*, some areas of the trench-region are not etched. Instead, as shown in Fig. 9-26,[42] regions of Si remain in the center of trench. This has the same effect as if the trench width was reduced, and the degree to which dishing occurs decreases with trench width. Consequently, the



Fig. 9-26 Dummy (b) and dishing support features of small (a) and large wafer (c) features can prevent dishing.[50] Reprinted by permission of Semiconductor International.

dummy structure method is also effective in reducing dishing effects to an acceptable degree. The CMP-removal rate over regions that are densely packed with active areas is also slower than over regions containing small isolated areas. Thus, to expose the nitride-CMP-stop-layer over such densely packed regions, the nitride over small isolated active-areas gets stripped, causing damage to Si in areas beneath the stripped nitride. As a result, dummy-active-regions must also be added to areas around isolated active regions. The dummy-structure technique, however, poses several difficulties. First, it imposes electrical, design, and space constraints, with the main problem being that several times as many structures must be added within a die over previous designs. This complicates the design and mask making process, and significantly raises defect levels. In addition, such dummy structures in trenches cannot be used for analog- and mixed-signal chips because the coupling capacitance over trenches becomes variable, and this disrupts the correct operation of analog circuitry.

In Procedure **3** used to reduce dishing during trench-CMP, a thin layer of silicon nitride is blanket-deposited onto the trench-oxide before it is polished-back with CMP (Fig. 9-27). A slurry with low selectivity between nitride and



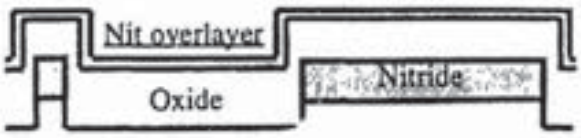


Fig. 9-27 *Nitride cap* approach to minimize dishing during CMP of the trench-fill oxide.

oxide is used for the CMP process, and thus the thin-nitride-layer on the top of the oxide covering the active areas (which are "up" regions) is polished away. This exposes the oxide under the nitride in the active-areas to the CMP mechanism. However, the surface of oxide that fills the trenches is recessed below the surface of the oxide on the active-regions (as these are "down" regions). Hence, the SiN on these recessed-regions is not polished away when the areas of nitride on the active-regions are removed. This SiN thus remains on the wafer to act as a polish-stop-layer when the trench-oxide-thickness is reduced by the CMP process to the degree that it is essentially planar with the surface of the active-regions. In this way dishing over wide trench regions is decreased.

In Procedure **4**, the thin nitride layer is again deposited on top of the trench oxide surface before CMP, as is done in Procedure **3**.[44] However, in Procedure **4**, this thin nitride is also patterned with a mask such that after this lithography step is completed, the nitride on the active area regions is exposed (Fig. 9-28). This nitride is then etched, so that it is no longer present on the active-areas of the chip. However, the thin nitride does remain covering the trench regions, where it acts as a polish-stop-layer during the oxide-planarization CMP-step. Thus, dishing of the oxide in the trenches is again suppressed.

Use of the *direct-CMP method* avoids the added complexity of the previous methods used to combat the trench-dishing problem. That is, in the direct-CMP-approach the trench dielectric is etched directly after deposition and densification without applying any of the above procedures. To permit the trench-oxide to be directly polished, the CMP process itself must exhibit reduced tendency to dishing (putting the burden the CMP tool, the slurry, and the overall CMP process).

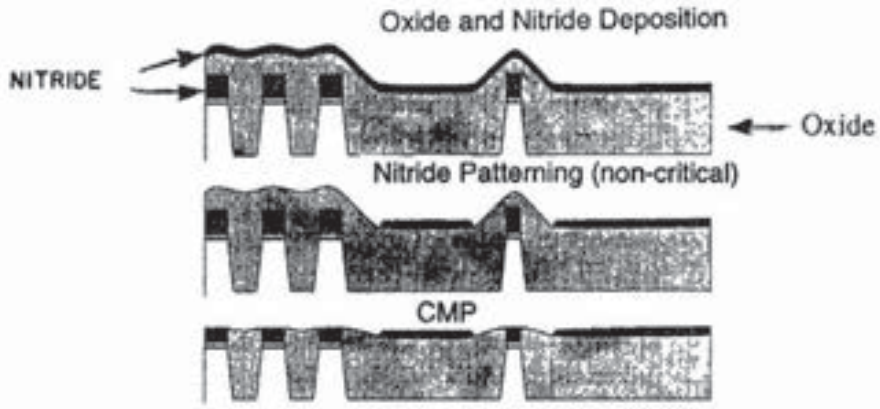


Fig. 9-28 *Patterned second-nitride layer* approach to minimize dishing during CMP of trench-fill oxide.

The three methods being pursued to allow direct-CMP are: **1**) the use of *linear polishing tools*; **2**) the use of *rotary polishing tools with a slurry that exhibits relatively low selectivity between oxide and nitride*; and **3**) the use of *rotary polishing tools with a fixed abrasive pad*. Linear polishers employ higher pad velocities (300–500-ft/min versus 34 ft/min for rotary pad speeds) and lower polishing pressures (3–5 psi versus 7 psi for rotary polishers). In the linear polishing process, a *highly selective slurry* (with cerium oxide particles dispersed in an aqueous suspension of material that inhibits silicon-nitride removal) is also employed.[45] This set of process conditions significantly reduces dishing effects (maximum oxide removal across a die was 20-nm) and decreases the nitride removal over isolated active regions.[46,56] On the other hand, dishing is reportedly decreased when *rotary polishing tools* are used by selecting a different set of process conditions and CMP consumables. That is, decreased dishing effects are obtained by using a harder polishing pad, a relatively non-selective slurry, efficient endpoint detection method, and an optimized CMP process.[47,48] The use of a fixed-abrasive pad for the direct-CMP method of STI trenches has also reportedly shown good results.[49]

The degree to which a CMP process used to polish STI trenches causes dishing can be evaluated with the help of a special STI density- and pitch-pattern test-wafer developed by MIT.[55] These test-wafers have 4800-Å-deep trenches, 8000-Å-TEOS oxide deposited over 1960-Å of silicon nitride and 150-Å of thermal liner-oxide. Process engineers can use this test wafer to measure the effects of trench dishing and nitride erosion.

9.5.5 Corner Engineering and Its Effects on MOSFET TURN-OFF Characteristics of MOSFETs Fabricated with STI

As we described in Section 9.1, the relationship of the elevation of the isolation dielectric surface to that of the active region of the MOSFET can impact the TURN-OFF behavior of the MOSFET. Specifically, if there is a recess of the isolation-region dielectric with respect to the active-region-edge at the top corner of the trench-edge, the polysilicon gate will wrap around the corner as shown in Figs. 9-3 and 9-29.[51] This can cause the leakage current of the MOSFET in the subthreshold region of operation to increase above the value it would exhibit if the polysilicon and the isolation were at the same elevation (and there then would be no poly wrap-around).

In the sequence that described the baseline-STI structure we indicated that there are three HF-deglaze steps that typically occur in such a process flow. The effect of the latter two (see Steps **17** and **20**) is to cause the dielectric in the trench to be etched at the same time the pad oxide and sacrificial oxides are etched (in Steps **16** and **19**, respectively). The effect on the topography of the dielectric at the top corner of the trench is shown in Fig. 9-30, in which it is seen that a *divot* is formed in the trench oxide at the active-area edge.[52] This divot can have four related and detrimental effects on the active MOSFET. First, it causes poly wrap-around, that may lead to increased leakage of the MOSFET operating in the subthreshold region. As described in the baseline-STI process, however, if sufficient rounding of the top corner is implemented, the enhanced subthreshold leakage (as evidenced by the "double-hump" or "kink" in the MOSFET TURN-OFF characteristics) can be suppressed. Second, the wrap-around causes the MOSFET to behave like a MOSFET which has an isolation region recessed

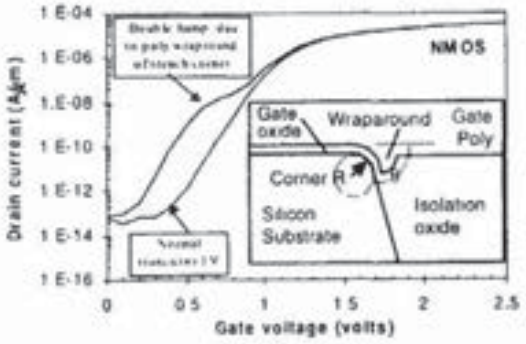


Fig. 9-29 "Kink" in MOSFET I_D-V_{GS} characteristic due to parasitic corner conduction caused by poly wraparound.

below the silicon surface. This enhances the so-called reverse-narrow-width effect, in which the threshold voltage decreases with MOSFET channel width, (see Fig. 9-31).[53] Note that this figure also shows that corner rounding reduces this effect, even if some poly wraparound occurs. Third, the threshold voltage of the MOSFET decreases as the depth of the divot is increased (Fig. 9-30). These effects produce unacceptably large device threshold voltage variations (± 200 mV). Fourth, the rounded edge encroaches into the active region of the device (i.e., under the nitride edge). This encroachment reduces the electrical width of the MOSFET. As the 0.1-μm technology node is approached, this width-reduction can represent a high percentage of the width of a minimum-sized MOSFET.

One reported technique that is used to combat this problem involves implementing a new corner rounding process that is accomplished after the nitride hard mask is opened (but before the silicon trench etch is performed). That is, a new lithographic technique from Clariant-AZ termed RELACS,[54] causes the trench width to be slightly reduced at the time the silicon nitride layer is opened. This is accomplished by applying a material to the resist pattern after it has been exposed and developed, and this material remains attached as a thin layer to the sidewalls of the resist pattern. This makes the openings of the resist pattern that define the silicon-nitride patterns slightly smaller (Fig. 9-32a). After the nitride layer has been opened, a short etch of the silicon trench surface undercuts the nitride edge, leaving a rounded top corner on the Si trench corner surface. The remainder of the trench is etched with an anisotropic etch. The overall trench profile after these etches is shown in Figs. 9-32b and 9-32c (here with the nitride left in place to show the location of the rounding with respect to the active area defined by the silicon nitride edge). The top corner rounding is clearly visible.

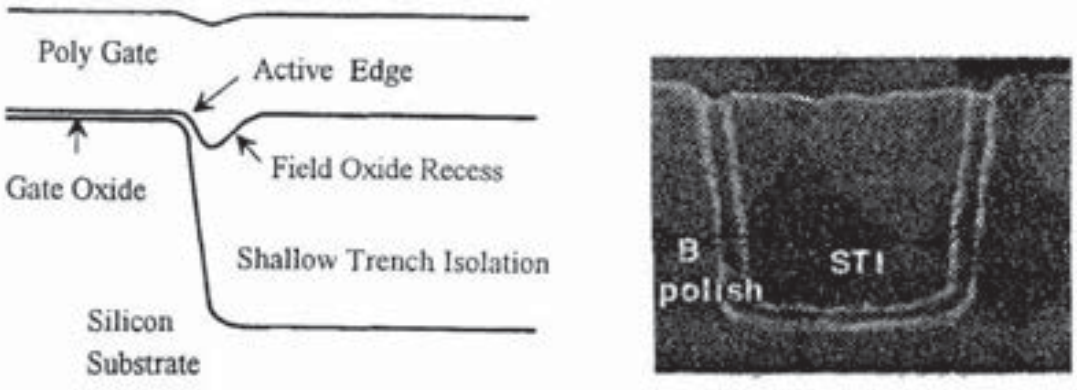


Fig. 9-30 Field-oxide recess (or *divot*) formed at the active edge of an STI structure.

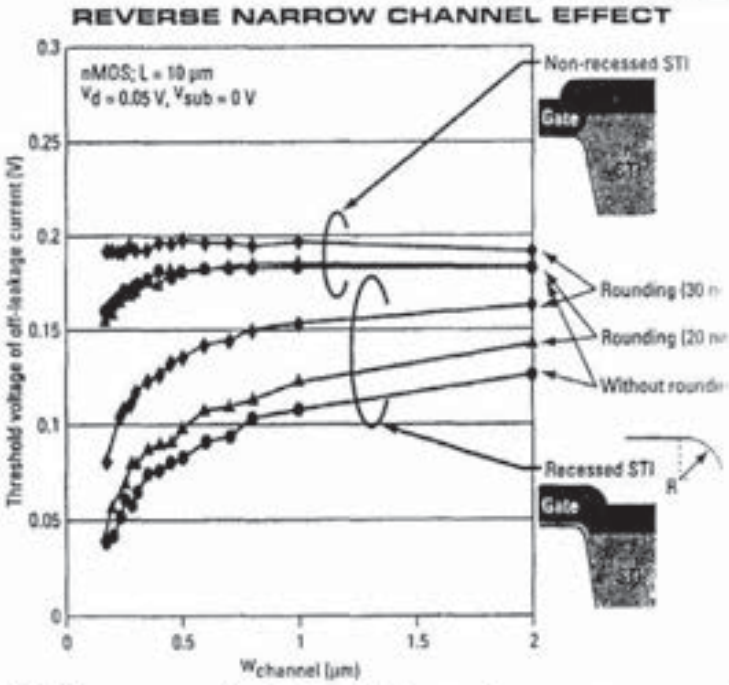


Fig. 9-31 Corner rounding helps eliminate the reverse narrow channel effect that causes subthreshold leakage. Recess of the isolation region below the silicon surface exacerbates the problem.[26] Reprinted with permission of Semiconductor International.

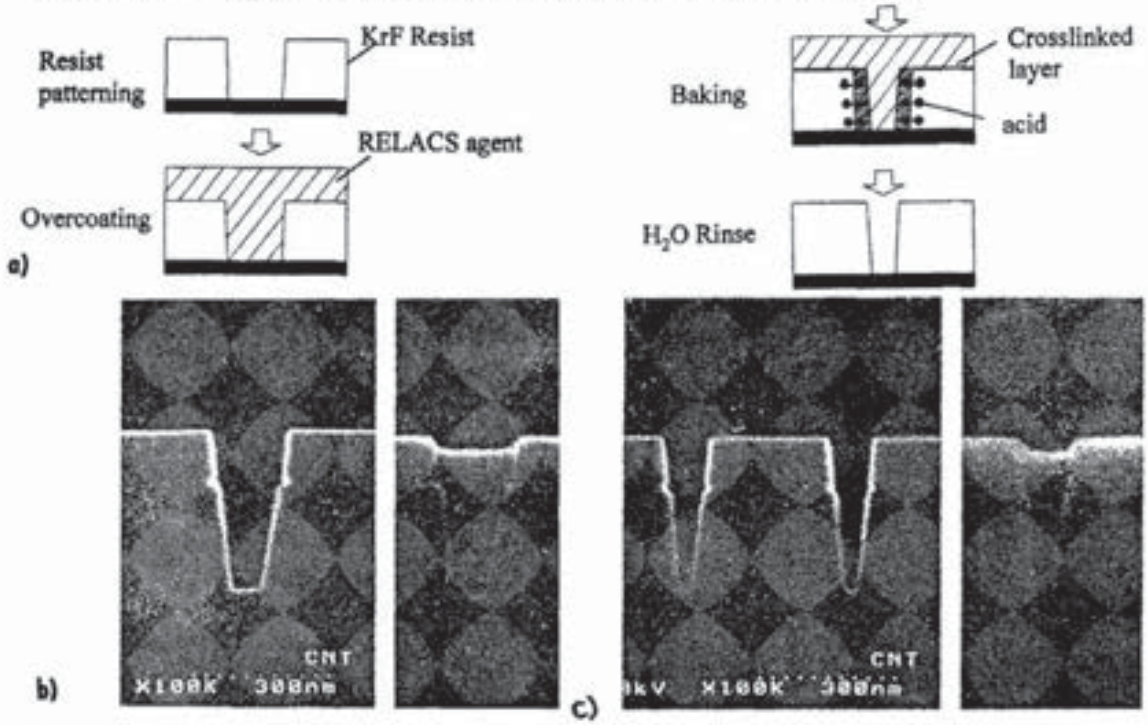


Fig. 9-32 a) Process flow of RELACS. (b & c) Top corner rounding provided during etch in a TCP system. An *in-situ* hard mask/rounding/trench etch scheme was used. The top corner is rounded to ~15 nm radius. b) 180-nm-wide trenches as etched and after fill and CMP; c) 100-nmwide trenches as etched and after fill and CMP.[54]

By providing the top-corner-rounding during the etch, the encroachment under the nitride edge by the rounded corner produced in the baseline-STI sequence is eliminated, and the severity of the poly wrap-around phenomenon is also reduced. Furthermore, the corner is rounded without encroachment. Another benefit of this technique is that less variability in the threshold-voltage value is observed from-wafer-to-wafer. The reverse narrow-channel effect becomes more severe if the trench-oxide encroaches beneath the silicon-nitride edge.

REFERENCES

1. E. DuBois *et al.*, "A Study of the Electrical Performances of Isolation Structures," *IEEE Trans. Electron Dev,*. **ED-37**(6), p. 1477, June 1990.
2. K. Kurosawa, T. Shibata, and H. Iizuka, "A New Bird's-Beak-Free Field Isolation Technology for VLSI devices," *IEDM Tech. Dig.* 1981, p. 384.
3. T. Shibata *et al*, *IEDM Tech. Dig.* 1983, p. 27.
4. D. C. Cole *et al.*, "The Use of Simulation in Semiconductor Technology Development," *Solid State Electronics*, **33**, No. 6, p. 591, 1990.
5. D. Foty, J. Mandelman, and T. Furukawa, "Behavior of an NMOS Trench-Isolated Corner Parasitic Device at Low Temperature," *Electrochemical Soc. Mtg. Abs.*, Fall, 1989, p. 295.
6. T. Furukawa, J. Mandelman, "Process and Device Simulation of Trench Isolation Corner Parasitic Device," *Electrochem. Soc. Mtg. Abs.*, Fall 1988, p. 329.
7. N. Shigyo *et al.*, "Three-Dimensional Analysis of Subthreshold Swing and Transconductance for Fully Recessed Oxide (Trench) Isolated 0.25-μm-Width MOSFETs," *IEEE Trans. Electron Dev.*, **ED-35**, July 1988, p. 944.
8. A. Bryant *et al.*, "The Current-Carrying Corner Inherent to Trench Isolation," *IEEE Electron Dev. Letts.*, **EDL-14**, August 1993, p. 412.
9. T. Iizuka *et al.*, "Double Threshold MOSFETs in Bird's-Beak Free Structures," *Tech. Dig. IEDM* 1981, p. 384.
10. K. Ohe *et al.*, "Narrow-Width Effects of Shallow-Trench-Isolated CMOS with n^+ Polysi Gate," *IEEE Trans. Electron Dev,*. **ED-36**, June 1989, p. 1110.
11. J.Y. Chen, "CMOS - The Emerging Technology," *IEEE Circuits and Devices Magazine,* March 1986, p. 16.
12. J.R. Brews, "The Submicron MOSFET," in *High-Speed Semiconductor Devices,* Ed. S.M. Sze, John Wiley & Sons, New York, 1990, p. 183.
13. K.M. Cham *et al.*, "Characterization and Modeling of the Trench Surface Inversion Problem for the Trench Isolated CMOS technology," *Tech. Dig. IEDM* 1983, p. 23.
14. *ibid. Ref. 12.*
15. *ibid. Ref. 12, p. 185.*

16. G. Fuse, "A New Isolation Method with Boron-Implanted Sidewalls for Controlling Narrow-Width Effect," *IEEE Trans. Electron Dev.*, **ED-34**(2), p. 356, Feb. 1987.

17. B. Davari *et al.*, "A Variable-Size Shallow Trench Isolation Technology with Diffused Sidewall Doping for Submicron CMOS," *IEDM Tech. Dig.*, 1988 p. 92.

18. S. Ambadi *et al.*, "Impacting Device Performance and Yield Through Sacrificial Oxidation Improvements," *Mtg. Abs. Electrochem. Soc.*, Fall 2000, Abs. 713.

19. S. Lassig *et al.*, "An Integrated etch Approach as STI Evolves for the 100-nm Regime," *Solid State Technology*, July 2000, p. 157.

20. L. Peters, "Gaining Control Over STI Processes," *Semiconductor International*, March 2000, p. 74.

21. A. Ambadi *et al.*, "Development of Shallow Trenches with Round Bottoms and Factors Affecting Etch Profile and Uniformity," *Mtg. Abs. Electrochem.. Soc.*, Fall 1999, Abs. 1312.

22. N. Nandakumar *et al.*, "Shallow Trench Isolation for Advanced ULSI CMOS Technologies," *Tech.. Dig. IEDM*, 1998, p. 133.

23. N. Balasubramanian, "Gate Oxide Thinning in MOS Structures with Shallow Trench Isolation," *Mtg. Abs. Electrochem. Soc.*, Spring 2000, Abs. 454.

24. *ibid., Ref. 20.*

25. *ibid., Ref. 19.*

26. L. Peters, "Choices and Challenges for Shallow Trench Isolation." *Semiconductor International*, April 1999, p. 69.

27. *ibid., Ref. 22.*

28. A. Chatterjee, *J. Vac. Sci. Technol.*, 1997, p. 1936.

29. C.P. Chang *et al.*, "A Highly Manufacturable Corner Rounding Solution for 0.18-μm Shallow Trench Isolation," *Tech. Dig. IEDM*, 1997, p. 661.

30. A. Chatterjee *et al.*, "Shallow Trench Isolation Using LOCOS Edge for Preventing Corner Effects for 0.25/0.18-μm CMOS Technology & Beyond," *Tech. Dig. IEDM*, 1996 p. 829.

31. T. Sato *et al.*, "Trench Transformation Technology using Hydrogen annealing for Realizing Highly reliable Device Structure with Thin Dielectric Films," *Dig. Symp. VLSI Tech.*, 1998, p. 206.

32. T. Sato, "Novel Corner-Rounding Process for Shallow Trench Isolation Utilizing Micro-Structure Transformation of Silicon," *Tech. Dig. IEDM*, 1998, p. 137.

33. S. Nag *et al.*, "Comparative Evaluation of Gap-Fill Dielectrics in Shallow Trench Isolation for Sub-0.25 μm Technologies," *Tech. Dig. IEDM*, 1996, p. 841.

34. L.-Q. Xia *et al.*, "High Aspect Ratio Trench filling Using Two-Step Subatmospheric Chemical Vapor Deposited Borophosphosilicate Glass for <0.18-μm *Device Application," J. Electrochem. Soc.*, May 1999 p. 1884,

35. L.-Q. Xia *et al.*, "High Temperature Subatmospheric CVD Deposited Undoped Silicate Glass," *J. Electrochem. Soc.*, March 1999 p. 1181.

36. S.S. Roth *et al.*, "PELOX Integrated PBL," *IEEE Trans. Semi. Manf.*, **6**, August, 1993, p. 246.

37. *ibid., Ref. 22.*

38. *ibid. Ref. 26.*

39. G. Badenes *et al*, "A New Dummy-free Shallow Trench Isolation Concept for Mixed-Signal Applications," *J. Electrochem. Soc.*, October 2000, p. 3827.

40. *ibid., Ref. 22.*

41. *ibid., Ref. 26.*

42. P. Tsui *et al.*, "A Manufacturable and Modular 0.25-μm CMOS Platform Technology," *Dig. Symp. VLSI Technology,* 1998, p. 152.

43. P. Singer, "The Future of Dielectric CVD: High-Density Plasmas?" *Semiconductor International*, July 1997, p. 127.

44. *ibid. Ref. 39.*

45. J.Y. Kim. "Novel STI CMP Process Using High Selective Ceria Slurry with Thin Nitride stopper," *Mtg. Abs. Electrochem. Soc.*, Spring 2000, Abs. 379.

46. B. Withers *et al.*, "Wide Margin CMP for STI," *Solid State Technology*, July 1998, p. 173.

47. *ibid. Ref. 26.*

48. *ibid. Ref. 26.*

49. J. Schlueter, "Trench Warfare: CMP and Shallow Trench Isolation," *Semiconductor International,* October 1999, p. 123.

50. R. DeJule, "CMP Challenges Below a Quarter Micron," *Semiconductor International,* November 1997, p. 54.

51. *ibid. Ref. 49.*

52. *ibid., Ref. 19.*

53. *ibid., Ref. 48.*

54. T. Toyoshima *et al.*, "0.1-μm-Level Contact-Hole Pattern Formation with KrF Lithography by Resolution Enhancement Lithography Assisted by Chemical Shrink (RELACS)," *Tech. Dig. IEDM,* 1998, p. 333.

55. Massachusetts Institute of Technology.

56. E. Zhao and C. Shan Xu, "Direct CMP for STI," *Semiconductor International,* June 2001, p. 145.

Chapter 10

SILICON-GERMANIUM (Si:Ge) TECHNOLOGY FOR HIGH-PERFORMANCE TELECOMMUNICATIONS ICs

Silicon-Germanium (Si:Ge) BiCMOS technology has become an important technology with many new and exciting product applications. Si:Ge-based heterojunction-bipolar-transistors (HBTs) are now found in a wide variety of technology offerings, and are incorporated in the product roadmaps of many telecommunications companies. In this chapter we describe the process technology issues of fabricating Si:Ge-HBTs and Si:Ge-BiCMOS ICs.

10.1 INTRODUCTION TO TELECOMMUNICATIONS IC APPLICATIONS

There is a rapidly growing market for wide-bandwidth and wireless telecommunications. That is, consumer demand for data from the Internet, for digital audio and video, and for video-on-demand is on the rise. In general, data is delivered to the consumer in two ways: *guided-wave propagation* and *free-space transmission*. Guided-wave propagation occurs along optical-fiber links and coaxial cables. Free-space transmission is the territory of radio, television broadcasting, satellite links, local multi-point distribution systems, and wireless local-area networks. (For example, such wireless applications as cellular phones currently operate in the 2.5-10-GHz range.) For free-space propagation, expensive installations are not necessarily required, but governments do regulate the use of bandwidth and frequency with a firm hand.

Since most of the *wireless* part of the frequency spectrum has been allocated, new or large bandwidth applications will be forced to resort to unallocated higher carrier frequencies. For example, a new wireless local-area network may have to operate at 60 GHz. Besides possible economy, a second attraction of higher frequencies is that they use smaller antennas. Because antenna efficiency is proportional to carrier frequency, receivers of higher carrier frequencies need only compact antennas. For instance, an antenna just 10-cm in length suffices for

collision-avoidance automotive radar at 77-GHz.

As for *guided-wave systems*, cable has been deployed, and once it is in place, bandwidth is relatively unregulated. In this kind of system (such as an optical-fiber data-link), the development of high-frequency components is mainly a response to the call for higher data rates. As optical-fiber trunk lines are more fully exploited, the next generation of communications channels will come to operate at 40, 80, or 100 Gb/s, well above the current 2.5-10 Gb/s. In fact, 100 Gb/s links could be in production as soon as 2010.

These data rates raise another obstacle. The signals must be received and processed at high speed and then eventually slowed down to match silicon CMOS processing speeds of, at best, 1.5-Gb/s. Such a receiving circuit requires complexity of at least the LSI level.

There is also a trend to higher operation frequencies in mobile communications and high-speed data-links. Future optical communications systems will also operate at over 10-Gb/sec and microwave/millimeter-wave systems will be used for mobile communication and intelligent traffic-control. The diverse range of applications in the broadband space includes *local multi-point distribution systems* (LMDS) that distribute information within major metropolitan centers via microwave transmission; *global-positioning systems* (GPS); *wide-* and *local-area networks*; *cellular base stations* and *handsets*,

Table 10-1 Performance Characteristic Comparison of GaAs and Si:Ge-HBTs

Characteristic	GaAs HBT	Si:Ge HBT
GAIN		
f_t, GHz	50	30
f_{max}, GHz	70	62
G_{max} @ 2 GHz, dB	20	23
NOISE		
F_{min} @ 2 GHz	1.5	0.5
G_{assoc}	> 20	> 0.5
1/f corner	1-10	0.1-1
POWER		
V_{be}, V	1.4	0.85
V_{knee}, V	0.5	0.25
FEATURE SIZE (μm)	2.0	0.5

including *cellular, 3rd-generation* (3G) *personal-communications systems* (PCS) for wireless telephony; *asynchronous transfer mode* (ATM) *switching and multiplexing for information transfer*; and the *synchronous optical network* (SONET) protocols and systems for optical communication.

In 1990, conventional wisdom predicted that devices built using silicon-based technology could operate only at high frequencies below 2-GHz. As noted above, there are many applications in telecommunications, radar, and data transmission that need higher frequencies. It was felt that all of these applications would require devices built using the more expensive III-V-based materials. In fact, since that time such wireless and high-performance (wide bandwidth) applications have turned increasingly to gallium arsenide (GaAs) and other III-V compound-semiconductor technologies to fabricate the high-frequency devices that such applications require. However, in the wireless personal communications systems that use GaAs devices, designers have had to use many technologies for one system (such as GaAs technology for the power amplifier [PA], low-noise amplifier [LNA] and other rf components), and CMOS-silicon-VLSI-technology for the baseband operations. To overcome this issue, in 1992 IBM decided to develop a competitive, commercially-acceptable analog technology for such applications based entirely on silicon (that is to say, strictly speaking, on silicon devices fabricated with some added *silicon-germanium* [Si:Ge] layers).[36]

The premise for this technical and commercial breakthrough was to transform standard Si into a heterojunction semiconductor material by doping the Si lattice with germanium (Ge). Such a film can be created by epitaxially depositing such a Si:Ge layer onto a Si substrate. This epitaxial Si:Ge layer can be used as the base region of a *heterojunction bipolar transistor* (HBT). Table 10-1 compares the performance of such Si:Ge-HBTs to those made with GaAs. It can be seen that the maximum operating frequency of operation of Si:Ge-HBTs is close to those made from GaAs.

When IBM decided to pursue this path, they realized that there were many process integration challenges that would have to be met. However, if the venture was successful, it would become possible for silicon-based devices to compete with III-V semiconductor performance while simultaneously maintaining the low-cost advantages traditionally associated with Si. Furthermore, the intent was to make the ultra-high-speed Si:Ge-HBTs fully compatible with existing CMOS technologies. This would enable ICs to be designed with a BiCMOS process that integrated these ultra-high-frequency (bipolar) HBTs with CMOS on the same chip. In this fashion, a technology would be created that combines the radio-frequency (rf) front-end circuits with CMOS signal-processing circuitry. Since

such a BiCMOS technology would offer both high-speed HBTs and CMOS on the same chip, it would be ideally suited for providing the high-performance and low-power characteristics required by wireless applications. In addition, it would be possible to manufacture the products on 200-mm wafers, with low defect densities, and in mature silicon-IC production facilities.

The overall approach to establishing such a technology was to develop a process in which the fabrication sequence of the Si:Ge-HBT became a module that could be integrated into a state-of-the-art CMOS process with no change to the CMOS flow or parameters. If this could be possible, it would represent an attractive route to cost-effective, single-chip solutions for the wireless market. That is, the cost of manufacturing Si:Ge-ICs is only marginally higher than for ICs in Si alone, while the fabrication of ICs made of GaAs or InP is considerably higher. In addition, such an approach would also pave the way for achieving high levels of integration not possible with GaAs technologies. One of the pioneers of Si:Ge technology, Bernard Meyerson of IBM, was asked in 1999 about the future of Si:Ge technology. He replied, "In the next 3-5 years, the entire world will move from global systems for mobile communications to code division multiple access (CDMA) with added focus on high-speed data-links."[1]

10.1.1 Basic Research Underpins IBM's Si:Ge Bid

The decision by IBM to launch an IC product-line for wide-bandwidth and telecommunications applications based on Si:Ge-technology was bolstered by more than a decade of research that was performed at the IBM Watson Research Center on Si:Ge. The original intention when this research was launched in the mid-1980's was to extend bipolar technology for mainframe computing applications. While IBM eventually abandoned bipolar technology in favor of CMOS for this application, it nevertheless became apparent that due to their high-frequency performance and low-noise characteristics, Si:Ge *heterojunction bipolar transistors* (Si:Ge-HBTs) were also ideally suited for telecommunications applications. Thus, IBM decided to continue to fund the Si:Ge technology program, with the new application in mind.

IBM published the first papers on Si:Ge devices in 1989. The devices described in these papers demonstrated much better performance than conventional silicon devices. Maximum clock speeds were 75-GHz compared to 25 GHz for the fastest silicon devices. This high speed could be traded off for improved linearity and lower power. That is, a Si:Ge device running at 20-GHz requires only 10-20% of the power needed to run a conventional silicon device at the same speed. This reduction in power is critical to the operation of battery-

operated systems like cell phones. Operating at lower power improves the linearity, lowers the amount of noise generated, and increases reliability.

By 2000 Si:Ge-based HBTs were available with a below-10-ps ECL-gate-delay, and a cutoff frequency of over 100 GHz. In fact, with these operating characteristics, it is expected that Si:Ge-HBTs will gradually replace GaAs devices in such key antenna interface circuits as low-noise amplifiers, switches, and power amplifiers. Indeed, the development of Si:Ge-technology has opened the entire frequency range to silicon-based devices. That is, it is now possible to build complex high-speed systems on a chip which combine complex logic arrays, high-speed bipolar devices, and passive components. These ICs offer unmatched performance yet can be built as routine products in conventional commercial production fabs. Furthermore, essentially any new process step developed for silicon can be incorporated into Si:Ge-technology.

The first Si:Ge commercial process was qualified for production at IBM in 1996. This BiCMOS process revolved around an HBT with a maximum available power-gain frequency of 65-GHz, together with various active and passive support elements that allowed designers to produce a wide array of high-performance analog ICs.[2] The BiCMOS process used a pre-existing ASIC CMOS process to merge large CMOS digital logic-functions with a high-performance HBT. A variety of products have since been introduced using ICs built with this process. For example, in the second half of 1999, Alcatel announced a 10-Gb/sec SONET transceiver built with IBM's Si:Ge-technology. By the year 2000, IBM had developed its third generation of Si:Ge-technology, which includes 0.4-μm and 0.25-μm BiCMOS.[3,4] A cross-sectional SEM of a 0.2-μm Si:Ge HBT is shown in Fig. 10-1.[5]

By 2000 a large number of companies had joined the rush to carry out research, development, or production on Si:Ge-HBTs. While the adoption of Si:Ge ICs by 2001 is still at an early stage, the overall market for such products was rapidly growing due to the lower-cost, high-performance opportunity created at the boundary of Si and GaAs.[6] At the beginning of 1999, total sales of Si:Ge ICs was believed to be only $3.7 million. But, it is predicted that by 2005 the market for Si:Ge chips alone will exceed $1 billion, and the systems into which they will be inserted will have sales of between $5-$10 billion. Increasingly, participants in the analog segment of the IC industry are developing proprietary process recipes to compete for market share in this lucrative business.

However, there are some barriers that have limited the growth of the market for this technology (although, as we shall note, this situation is changing):

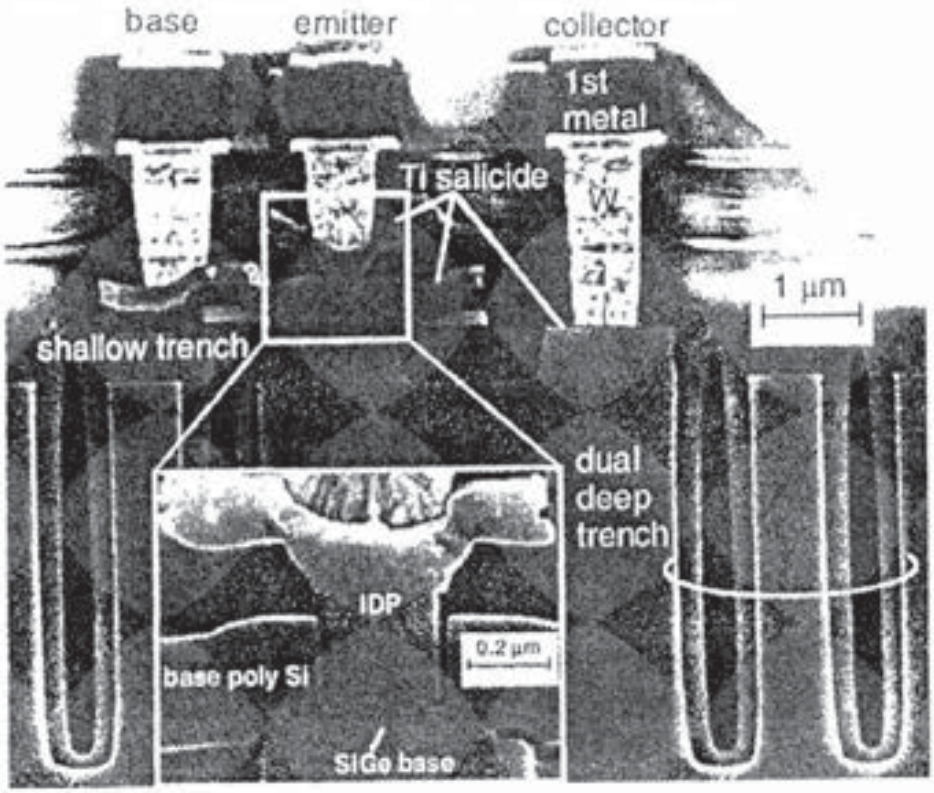


Fig. 10-1 SEM cross-sectional view of a 0.2-μm Si:Ge-HBT.[5]

1. The intellectual property is narrowly held by only a few organizations.

2. Until recently, relatively few manufacturers or foundries had developed the process-technology expertise for manufacturing such devices. However, by late-2000, most state-of-the-art BiCMOS foundries had announced Si/Si:Ge-HBTs as their baseline bipolar technologies.[37] For example, the Taiwanese foundry, TSMC, announced in early 2001 that they had licensed Si:Ge technology from Conexant.

3. The devices being made with Si:Ge are still primarily bipolar HBTs. The lack of commercially-available Si:Ge CMOS structures is therefore also an impediment. However, development of MOS-based Si:Ge devices has been described.[7]

10.2 INTRODUCTION TO HETEROJUNCTION BIPOLAR TRANSISTORS

As the vertical dimensions of the bipolar transistor were being scaled, it was predicted that some serious device-operational limits would be encountered, even if such innovations as the use polysilicon emitters were implemented (see Vol. 2, *Silicon Processing for the VLSI Era; Process Integration - Chapter 7 - Bipolar and BiCMOS Process Integration* for a detailed discussion on bipolar transistor device physics and the evolution of bipolar transistor technology). The following

scenario illustrates this point: When the base-width of the bipolar transistor is decreased, it's doping level must be increased in order to control punchthrough, and to maintain a sufficiently low-value of base resistance. Since the doping level in the emitter must remain high in order for a sufficient emitter-injection efficiency to be maintained, the emitter-base capacitance C_{EB} will also increase. The current density will thus need to be increased so that the propagation delay associated with this extra capacitance can be overcome. To handle the increased current density without suffering the Kirk effect, it will also be necessary to raise the doping in the collector. Consequently, even though such a device will have a higher f_T, it will have a narrower operating-current range. Furthermore, the increased doping levels will reduce the values of the emitter-base and collector base breakdown-voltages (BV_{EBO} and BV_{CEO}, respectively), to such a degree that it will not be possible to use the device in many practical applications.

One technique that can be used to overcome the above limits is to build bipolar heterojunction transistors (HBTs). An HBT is a bipolar transistor in which the emitter, base, or collector use materials with different energy bandgaps. Bipolar transistor speed is determined mostly by the time a charge-carrier (electron or hole) takes to pass through the base of the device and the time the junction capacitances take to charge up. Thinner base-region layers, higher charge-carrier velocities, and higher current densities all increase the frequency of operation, as do lower parasitic capacitances and lower device resistances.

Recognized metrics for the speed of transistors are the following:

1. *Current-gain cutoff frequency* (f_T): This is the frequency at which the current gain of the active-device falls to 1. The highest clock rate at which the technology based on these active devices can operate is related to f_T. The value of f_T is strongly influenced by minimum device dimensions in the direction of current flow and by device capacitances.

2. *Maximum frequency of oscillation* (f_{max}): This is the maximum frequency at which it is possible to make an oscillator out of a given device. It is also referred to as the *unity power gain cutoff frequency.* Unlike f_T (which is influenced by minimum device dimensions in the direction of current flow and device capacitances), the value of f_{max} is instead quite sensitive to parasitic elements external to the core of the device, such as the transistor contacts.

High values of both f_T and f_{max} are desirable for both digital and analog designs. While f_T and f_{max} values can be traded off against each other, balanced values are ideal for a general-purpose IC. Normally, small *gate lengths* in

MOSFETs and small *base widths* in bipolar transistors lead to high cutoff frequencies. If NMOSFETs are compared to Si:Ge bipolar transistors, it is found that NMOS devices with gate lengths of 0.18-μm have an f_T of about 70-GHz and an f_{max} of less than 40-GHz, while the most advanced Si:Ge-HBTs have cutoff frequencies in the 80-100-GHz range. It is predicted that Si:Ge-technology will be scaled to provide Si:Ge-HBTs with 150-GHz f_T-values by about 2005.[8] (Note that indium-phosphide [InP]-HBTs though not discussed here, have been fabricated with an f_T of 250-GHz, and an f_{max} over 1-THz [1000-GHz].)

10.3 HETEROJUNCTION BIPOLAR TRANSISTORS WITH LARGE-BANDGAP EMITTERS

One approach to fabricating HBTs is to build transistors with emitter materials whose band gaps are larger than that of the material used in the base. This technique appears promising because the emitter-injection-efficiency, γ, depends on the band-gap difference according to the equation:[9]

$$\gamma = (n_e \nu_e / p_b \nu_h) \exp(\Delta E_g / kT) \qquad (10.1)$$

where n_e is the emitter electron-concentration, ν_e is the effective velocity of electrons injected into the base, p_b is the base hole-concentration, ν_h is the corresponding hole velocity injected into the emitter, and ΔE_g is the difference in the band gaps of the emitter and base. From Fig. 10-2, it can be seen that the larger band gap in the emitter gives rise to a larger potential-energy barrier for holes going from the base to the emitter than for electrons going from the emitter to the base. Thus, if the band gap of the emitter in silicon devices is larger than that of the base, when ΔE_g is 0.2 eV or more, $\exp(\Delta E_g/kT)$ will exceed 2000. As a result, a sufficiently-high emitter-injection efficiency will be assured, no matter what doping levels exist in the base and emitter.

For example, doping levels in the base can be increased to above 10^{19} cm^{-3}, and in the emitter they can be dropped to 5×10^{17} $c\,m^{-3}$ at the emitter-base interface (Fig. 10-3a1). Under such conditions, many other device characteristics are also improved (among the improvements are a lower base resistance, a higher punchthrough voltage, a higher Early voltage, and a decreased E-B junction capacitance). However, for such devices to be successful in silicon technology, the technology must be compatible with advanced silicon bipolar processing.

Most approaches to heterojunction fabrication are based on growing an epitaxial layer on the Si substrate, such that the epi material has a different band gap than the Si, but also has a lattice constant that is at least approximately

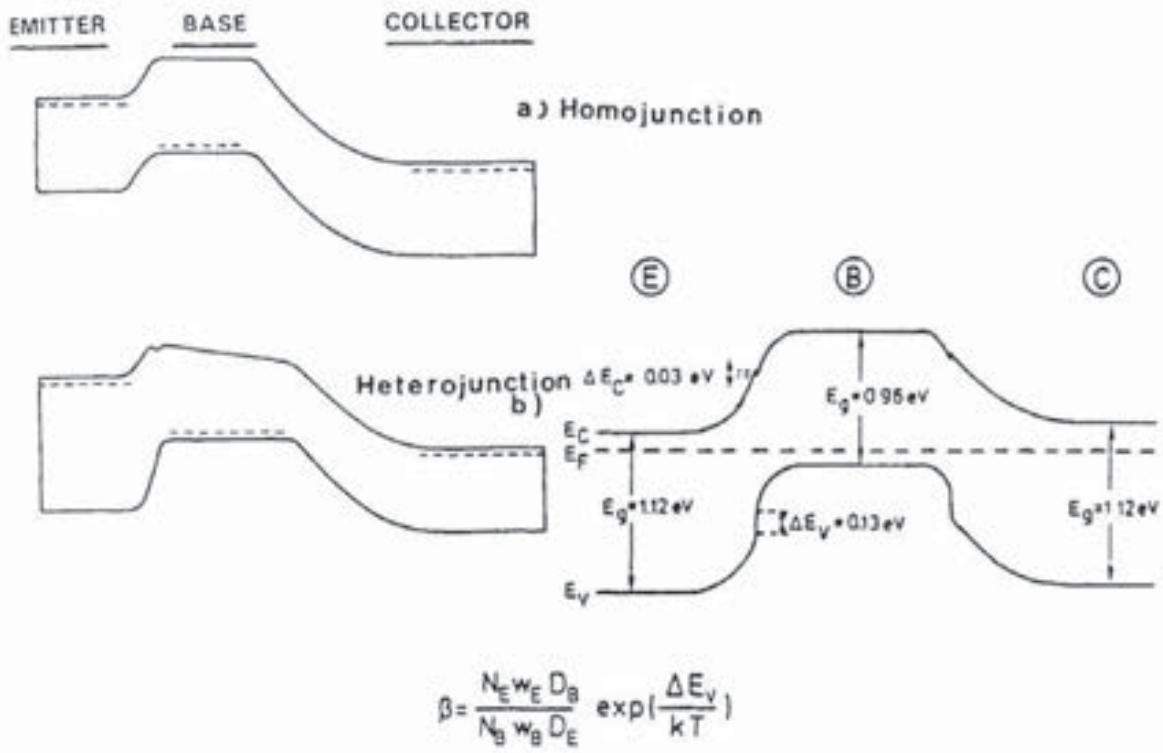


Fig. 10-2 Band diagrams of representative *npn:* a) homojunction, and; b) heterojunction bipolar transistors.[9] (© 1989 IEEE)

matched to the lattice constants of the Si substrate. If such deposited layers are sufficiently thin, the requirement for exact lattice-match is relaxed. That is, thin layers with an almost-matched lattice constant will deform so that they can accommodate to the lattice constant of the substrate. In this deformed condition they are either "coherently strained" or "pseudomorphic" layers. The latter are thermodynamically stable for thicknesses below a strain-dependent "critical" value. However, it is important that subsequent processing temperatures be restricted to prevent the stress of the layers from being relaxed through the formation of an array of misfit dislocations.

From the point of view of process integration, the most attractive approach to building devices with hetero-E-B junctions is to increase the band gap of the emitter (since this represents the only change from a conventional bipolar process). Several methods for implementing this approach have been attempted, including the use of the following materials as the wider band-gap emitters: GaP;[10] semi-insulating polycrystalline-silicon (SIPOS, E_g ~1.7 eV);[11] oxygen-

doped Si-epitaxial films (OXSEF, E_g ~1.7 eV);[12] β-SiC (E_g ~2.2 eV);[13] and phosphorus-doped hydrogenated micro-crystalline silicon (μc-Si, E_g ~1.7 eV).[14]

Despite an intense effort to develop such processes, no completely successful fabrication procedure using an emitter with a larger bandgap was demonstrated. Although numerous high-gain transistors have been reported, these devices exhibited other problems, including high-temperature requirements, high bulk and contact resistivities, high emitter resistances, or poor stability during the multilevel-interconnect fabrication steps. The most promising candidate of this group was the μc-Si emitter. Fujioka *et al.* reported achieving reduced emitter resistances with such structures and built ECL gate arrays with them using a non-self-aligned emitter structure.[14] Basic gate delays of 295-ps at 6.5-mW/gate were obtained.

10.4 HETEROJUNCTION BIPOLAR TRANSISTORS WITH SMALLER BANDGAP BASE REGIONS (SI:Ge BASE)

An alternative approach to using a *larger bandgap* material for the emitter is to use a material in the base of the transistor that has a *smaller bandgap* than that of

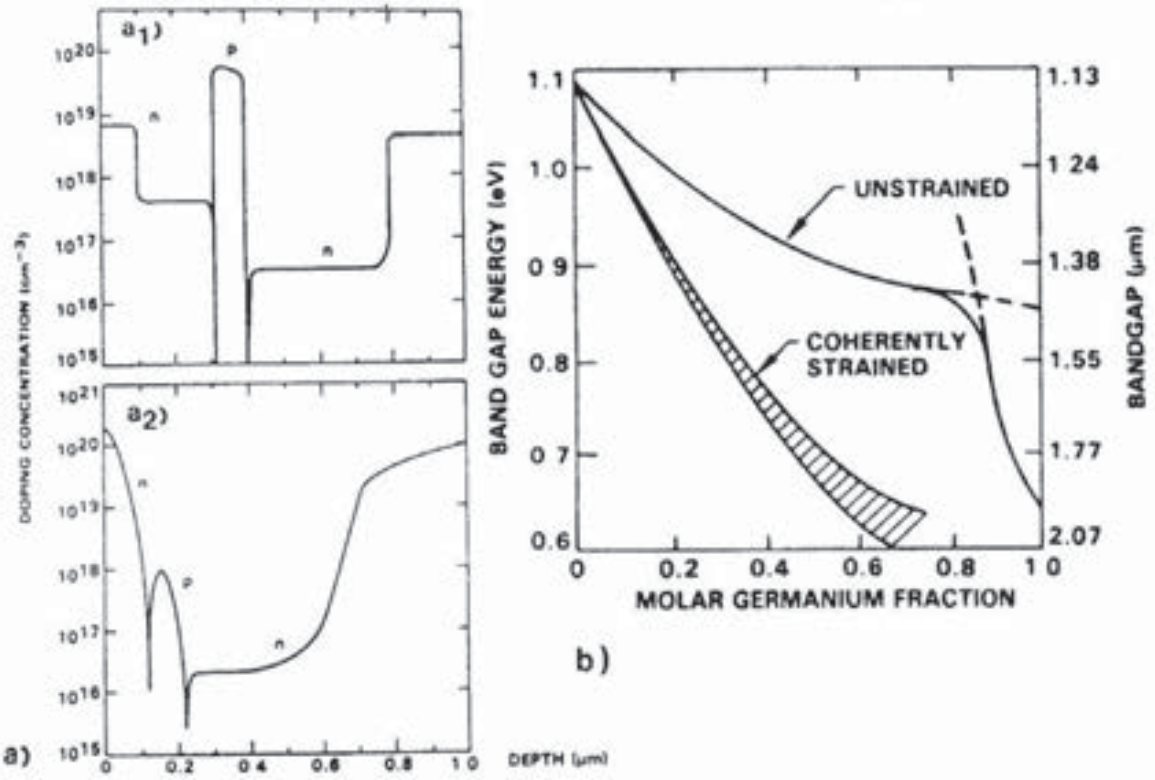


Fig. 10-3 a) Representative doping profiles for npn heterojunction bipolar transistors.[9] (© 1989 IEEE); b) Band-gap energy versus alloy composition for Si:Ge alloys.[19] (after People and Bean Copyright 1986 AIOP)

silicon. Layers of Si to which Ge has been added exhibit such reduced bandgaps (Fig. 10-3b). For example, pseudomorphic epitaxial Si_xGe_{1-x}-alloy layers of up to 100-nm in thickness can be formed with a Ge content of 20%, and the ΔE_g between the emitter and base in Si transistors fabricated with such base layers (i.e., *Si_xGe_{1-x}-base-HBTs*) can be as large as 0.1 eV.[15]

Pseudomorphic epitaxial Si_xGe_{1-x} layers can be deposited by means of molecular beam epitaxy (MBE) or limited-reaction-processing (LRP) CVD.[17] The most successful commercial CVD methods have involved ultra-high-vacuum-CVD (UHV-CVD),[16] or low-pressure CVD (LP-CVD). Note that appropriate levels of boron doping must also be incorporated into these films during the epi-growth process of the base-region in order to form the *p*-base of an *npn*-HBT. Only *npn-HBTs* are being pursued at present because *pnp*-HBT devices do not benefit substantially from the use of a Si:Ge base (due to a less favorable band-offset). We will describe the details of the UHV-CVD process used to form the Si:Ge-epitaxial-layers in the next section.

It should also be noted that to add a Si:Ge layer to a conventional Si device requires a special piece of semiconductor equipment (i.e., an MBE machine, or a special CVD reactor), which in 2000 cost approximately $2.5 million. However, the only other way to make a major performance improvement in a family of semiconductor devices is to shrink the geometry of the devices. Implementing a device shrink would typically require installing a whole new generation of process equipment and would cost more than $750 million!

HBTs with narrow-band-gap Si_xGe_{1-x} bases are able to achieve very high performances (e.g., f_T's as high as 100-GHz have been built with a 20%-Ge-concentration in a 100-nm base, a 0.2-μm-wide Si emitter, and a collector-current-density of $1x10^5$ A/cm^2).[18] Conversely, Si_xGe_{1-x}-base-HBTs offer the potential benefits of the same performance characteristics as those of conventional homojunction transistors, but with a relaxed profile (i.e., the base can be doped more heavily for a lower base-sheet resistance, or it can be made thicker to achieve lower doping levels with higher breakdown voltages).

The chief drawback of the reduced-band-gap-base approach is that it is more difficult to integrate into existing advanced bipolar processes. The metastable nature of the Si_xGe_{1-x} films places constraints on the Ge content, layer thickness, subsequent annealing cycles, and thermal oxidation. On the other hand, this method does not suffer from the other problems exhibited when wide-band-gap-emitter materials are used. In addition, to provide the ultrathin base profiles for future bipolar devices, the conventional ion-implantation approaches are becoming increasingly difficult to implement successfully, while the low-

temperature epitaxial processes show great promise. We will discuss the process technology and process integration issues of Si:Ge ICs in the following sections.

10.5 PROCESS TECHNOLOGY FOR FABRICATING Si:Ge-HBTS

As noted in the previous section, Si:Ge heterojunction bipolar transistors have the same general structure as homojunction bipolar transistors, with the exception that they contain a Si:Ge layer in place of a Si layer in the intrinsic-base-region of the transistor. This Si:Ge layer is formed by epitaxial deposition onto a Si substrate, and a polysilicon layer is deposited onto the Si:Ge base-layer to serve as the polysilicon emitter. Since only *npn*-HBTs are fabricated (i.e., as noted earlier, *pnp* HBTs do not exhibit substantial performance improvement over their homojunction *pnp* counterparts, so there is no advantage to manufacturing them), the *p*-type intrinsic base is fabricated by also incorporating boron into the Si:Ge layer as it is being grown.

The use of an epitaxial layer as a specific structure of a bipolar transistor (i.e., the base region), is a significant departure from past uses of epitaxial layers in the fabrication of Si ICs. That is, in conventional (i.e., non-Si:Ge) IC fabrication, epitaxial layers have been used to form a lightly-doped Si layer on a heavily-doped Si substrate. That is, the epitaxial layer becomes a part of the *substrate* in which the device is formed, but *does not act* as a *specific device structure* itself. The fact that the epitaxial Si:Ge layer in an HBT is an active-device-structure-layer allows one of the fundamental limitations of bipolar device fabrication to be overcome, namely the fact that the base region of modern homojunction bipolar transistors must be formed using an ion-implantation process.

The ion implantation process produces a base-doping-profile with a skewed Gaussian distribution (due to the Gaussian-distribution of the as-implanted dopants, plus channeling effects associated with the ion-implant process). By using an epitaxial process to form the base region, it is theoretically possible to grow a *uniformly-doped* layer with atomically abrupt boundaries, and all of this doping can be confined within the Si:Ge layer. In practice, such atomically abrupt boundaries are not possible, but one can create a boron-doping profile within the Si:Ge layer that is much more uniformly doped than are the base regions of bipolar transistors created with an ion-implantation process. Note that the Si:Ge epitaxial layers must be grown at temperatures low enough that the Si:Ge layers retain their pseudomorphic structure (500-550°C). Furthermore, by performing the epitaxial-deposition process at such low temperatures the outdiffusion of the dopant from the layer during the epitaxial growth process is essentially suppressed, keeping the boron well-confined within the Si:Ge layer.

A low-temperature-epitaxy (LTE) technique using ultra-high-vacuum chemical vapor deposition (UHV/CVD) in a multi-wafer reactor is the most mature process used to grow epitaxial layers of Si:Ge-alloys for use in HBTs.[21] It is reported that device-quality epitaxy can be achieved at a growth temperature of 550°C. This LTE technique originated in the early 1980's with Meyerson's work at IBM on a novel UHV/CVD reactor. The Si:Ge epitaxy process has two stages: **1**) preparation of the Si substrate surface so that an epitaxial layer can be grown on it at a low temperatures; and **2**) growth of the layer itself.

To enable growth of an epitaxial layer on a Si surface, the surface must be free of native oxide and other contaminants. To eliminate the need to use high temperatures it is necessary to ensure that the surface of the Si is free of such species prior to initiating the epitaxial-growth-stage of the process. A simple wet chemical procedure is carried out to remove any such native oxide from the Si surface and at the same time to passivate the wafer surface with hydrogen. This process entails etching the wafer for 10-15-sec in dilute (10:1) H_2O/HF. The layer of hydrogen adsorbed on the Si surface during this wet etch acts to retard the growth of native oxide on the Si surface. That is, after the 10-sec HF dip, there is a passivation layer consisting of hydrogen-terminated bonds across the silicon surface. The presence of the hydrogen reduces the reactivity of the growth interface by approximately 13 orders of magnitude from that of a bare-silicon surface with respect to its oxidation rate in ambient air. Since the wafer surfaces are oxide free and passivated as-pulled from the HF dip, the wafers may be loaded directly into the growth chamber (Fig. 10-4) after this HF-last step. This

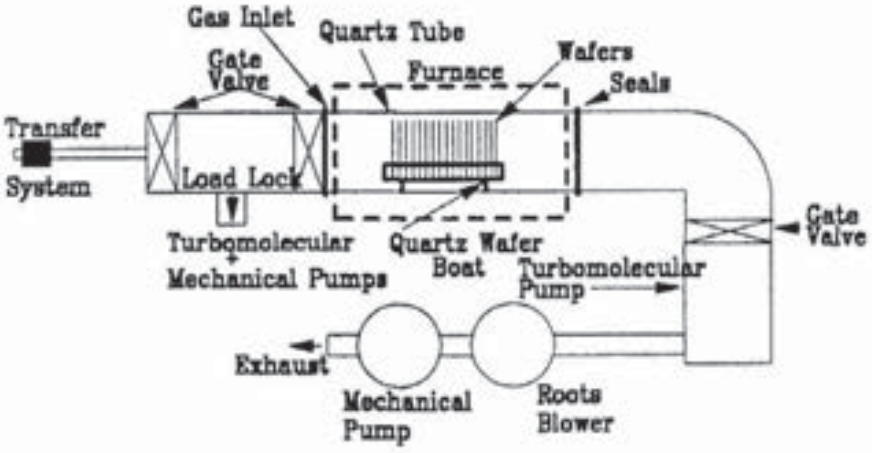


Fig. 10-4 Cross-section schematic of a UHV/CVD reactor.[15]

surface-preparation step has been shown to reliably keep the Si surface free of native oxide until the growth process can be begun (provided the temperature/time history of the load-sequence prevents hydrogen desorption until growth is initiated). Note that bare-silicon wafers removed from an HF bath are hydrophobic (i.e., they do not wet). Thus, the surface readily de-wets after being pulled from the HF. This makes it easy to eliminate residual HF from such a surface, allowing this removal process to be automated.

The batch of wafers (typically 25) is introduced into the loadlock of the UHV-CVD tool. After a pumpdown to 10^{-6} torr, the wafers are transferred into the growth chamber under an H_2 flow and growth is commenced immediately. The growth chamber is pumped down to a base pressure of 10^{-9} torr (UHV), and the precursor gases (silane, germane, and diborane) plus hydrogen are introduced. Films are deposited at temperatures in the range of 500-600°C. During deposition the pressure is maintained at 10^{-3} torr. The growth rate during this process is only about 0.4-4.0-nm/min. This very-low growth rate provides precise control of the layer thickness (on the order of 1-2 atomic layers in this instance), doping concentration, and formation of a pseudomorphic strained SiGe layer. Versions of this system are available for processing 200-mm wafers (based on IBM designs), from Unaxis (formerly Balzers Process Systems) and CVD Equipment Corp (Fig. 10-5). To ensure these tools are robust and productive in a manufacturing environment, good vacuum practice and contamination-prevention methods must be incorporated when designing them. Note that these reactors are batch systems that process up to 25 wafers at a time. The industry, however, is moving toward single-wafer processing, and single-wafer Si:Ge deposition tools are thus also available.

High-quality Si- and SiGe-epitaxial-films can be formed at such low temperatures for the following reason.[36] The hydrogen-terminated silicon wafer surface desorbs its hydrogen at temperatures between 500-600°C. The silane (SiH_4) that is fed into the reactor chamber and that adsorbs on the wafer surface also decomposes at these temperatures. Of the products of this decomposition, silicon incorporates into the crystal, and the hydrogen remains adsorbed on the silicon surface, replenishing the hydrogen that was initially there. Thus, at these temperatures the silicon-epitaxial-layer grows while maintaining a hydrogen-passivated surface. Thus, the surface is protected from ambient oxygen contamination, allowing high quality epitaxial layers to form. (Note that silicon can undergo solid-phase-recrystallization at reasonable rates in the range of 500-700°C). However, at higher temperatures (> 650°C), the hydrogen passivating film is removed instantaneously, allowing even low residual oxygen content in

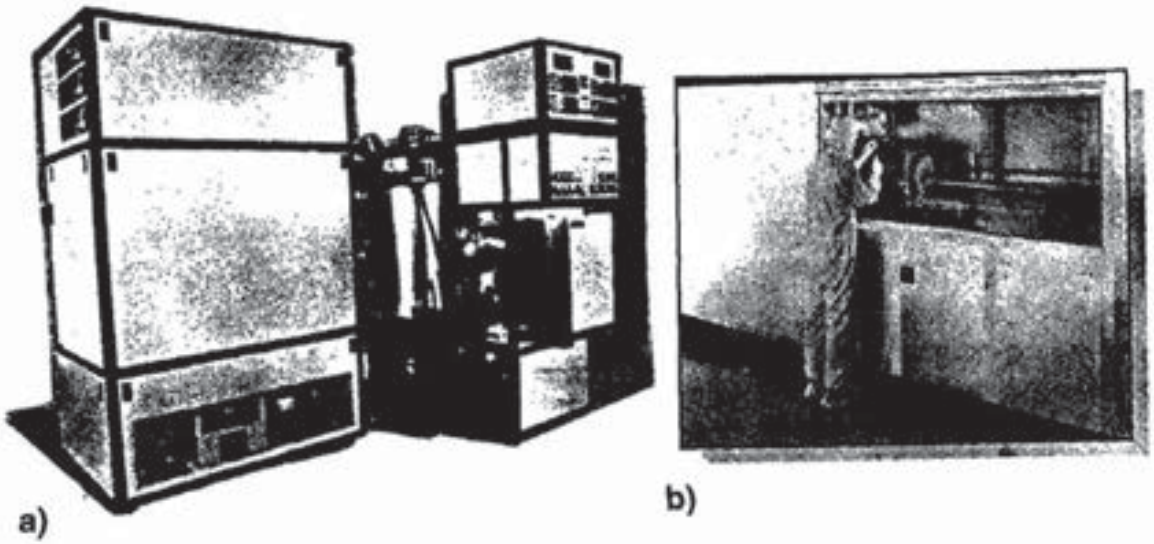


Fig. 10-5 a) Commercial UHV/CVD Si"Ge deposition system. Courtesy of CVD Systems; b) Commercial Low-energy plasma-enhanced CVD system for Si:Ge deposition. Courtesy of Unaxis Ltd.

the growth ambient to cause the formation of native oxide on the silicon surface. This causes high defect densities in the epitaxial film. As a result, to obtain high-quality films when silicon wafers are heated to temperatures greater than 650°C (initially hydrogen-passivated or not), a high-temperature pre-treatment step just prior to epitaxial growth is needed to remove the now-present native oxide.

Single-wafer, ultra-high-vacuum, rapid-thermal-CVD (UHV/RTCVD) has been reported as a means for achieving LTE.[22] An advantage of this single-wafer technology is that the process can be initiated and terminated by rapid changes in the substrate temperature, rather than by changing gases. Figure 10-6 shows a schematic of such a system.[37]

Conventional single-wafer epi reactors can also be operated with an LP-CVD process at low temperatures (about 600°C) to produce strained-layer Si:Ge-alloys at a deposition-rate of about 10-nm/min. Note that in single-wafer, load-locked, RTP tools, the surface-cleaning of the Si-substrate prior to the Si:Ge-epitaxial-growth stage is carried out by a high-temperature RTP step which sublimates SiO present on the wafer surface (instead of relying on an HF-etch as is done in the batch-UHV/CVD process). It must be rigorously ensured that the ambient during both the cleaning stage and the growth stage of the epitaxial-growth process in the single-wafer reactor is kept completely free of oxygen and moisture. This is accomplished by using a loadlock, point-of-use gas-purification, and ultra-clean gas handling. Two vendors that supply single-wafer reactors in which Si:Ge can be deposited by LP-CVD are ASM International[23,24] and Applied Materials.[39,40]

A substantial number of these reactors are reportedly being sold to the industry.[25] ASMI claims they have 75% of the installed Si:Ge-reactors in 2000 (ASMI Epsilon E2000®).

Etching the Si:Ge layers also presents some challenges. While Si:Ge can be etched with plasma processes that use chemistries similar to those used to etch Si, the fine features and profiles required to maintain high-frequency performance, place increased demands on the processes. Novel chemistries can be an aid, and endpoint-detection plays a vital role. Figure 10-7 shows a super-self-aligned selectively-grown-epitaxial Si:Ge-base transistor.

The use of *selective-epitaxial-growth* (SEG) to grow the Si:Ge-base of the HBT has been pursued to achieve higher f_T and f_{max} values. For example, a double-polysilicon self-aligned-HBT with a low-resistivity Si:Ge-base was fabricated having a 88-GHz f_T-value and a 78-GHz f_{max}-value. Si:Ge-HBTs with such performance characteristics would make then suitable for wireless communications up to 10-GHz and high-speed data-links at 40-Gbit/sec and above.[26] Another report described how such a Si:Ge-SEG-process was used to build a 0.2-μm, self-aligned Si:Ge-HBT that was then integrated into an SOI BiCMOS technology. The base Si:Ge layer in this HBT was deposited by UHV/CVD. The HBT in this report had an f_{max} of 180-GHz.[27]

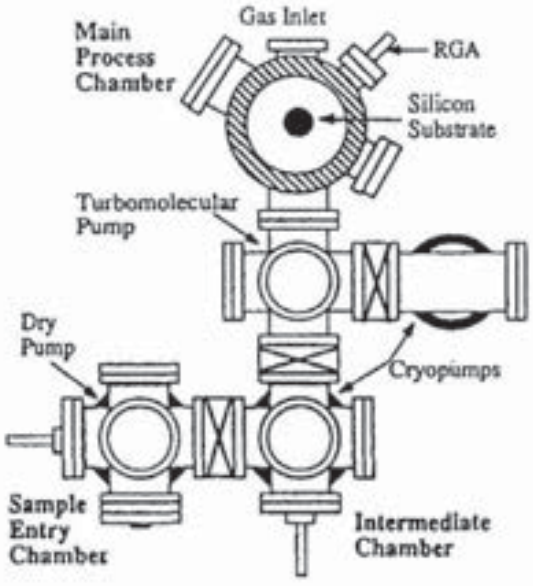


Fig. 10-6 Schematic of a UHV/RTCVD epitaxial system.[22] Reprinted with permission of the American Physical Society.

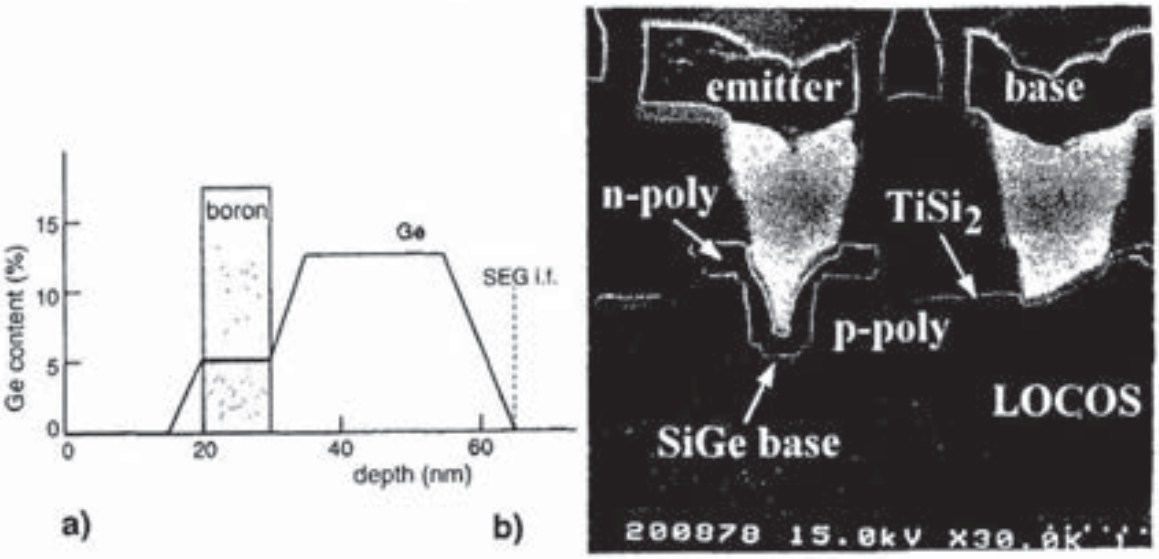


Fig. 10-7 a) Schematic of the Ge profile in the Si/Si:Ge layer of a Si:Ge-HBT;[27] b) SEM cross-section of the emitter-base complex of a 0.2-μm Si:Ge-HBT.[26]

10.6 PROCESS INTEGRATION ISSUES FOR BiCMOS TECHNOLOGY WITH Si:Ge-HBTs

The previous section dealt with the process technology of forming epitaxial layers of Si:Ge on Si-substrates. Here we consider the issues involved with integrating such an epitaxial-growth-step into a CMOS process so that HBT structures can be fabricated that are integrated on the same chip as CMOS circuitry. That is, a BiCMOS process is implemented instead of a CMOS process. The goal of creating such a BiCMOS technology is to end up with a process that results in no change (or minimal change) to the standard-CMOS flow, and that the electrical characteristics of the CMOS in a BiCMOS structure matches the baseline-CMOS as closely as possible (and conversely, that the CMOS process steps do not degrade the performance-characteristics of the HBT).

As noted earlier, basic physics explains that when a Si:Ge-lattice is formed, the atomic size difference between Si and Ge atoms necessarily results in the formation of a layer that will create a lattice-mismatch if the layer is formed on a Si substrate acting as a growth template. This lattice-mismatch must be dealt with when integrating this layer with the overall fabrication sequence. While a certain level of strain is required for the bandgap-reduction of the Si:Ge-layer (and this bandgap-reduction is a driving force behind HBT performance improvement), the strain must be kept low enough to avoid a generation of lattice defects and unstable Si:Ge films. This limitation has been extensively studied and the Matthews and Blakeslee curves relate film-thickness to strain in a Si/Si:Ge

system.[34] To allow the Si:Ge-layer to be integrated with the remainder of the processing steps used to fabricate standard CMOS, this implies that the Si:Ge-strained-layer must be able to tolerate total emitter-thermal-cycles such as an 850°C furnace-anneal followed by a short 900-1000°C rapid-thermal-anneal.[28]

10.7 HIGH-QUALITY Si:Ge EPITAXIAL FILM GROWTH ON PATTERNED WAFERS

One of the most significant issues facing the integration of an HBT device into a CMOS process is that of being able to deposit high-quality epitaxial films onto wafers containing pre-existing oxide-isolation-patterns. The use of wafers patterned prior to epitaxial deposition is the process factor that separates this approach from the approach that is used to fabricate discrete HBTs or mesa bipolar transistors that contain a Si:Ge-epitaxial- base-region. The problem is that the parts of the wafer covered by oxide are hydrophyillic (these surfaces *do* wet). The wetted surface makes it difficult to remove residual-HF without a DI-water rinse-and-dry step. But, such a water rinse would degrade the passivation quality of the hydrogen-termination layer present on the regions of the wafer where silicon is exposed (through hydrolysis of less stable F-terminated surface and/or kink sites which cannot be ideally passivated).

A pragmatic two-step solution when depositing Si:Ge-films in the UHV/CVD reactor is to first assure that the entire exposed wafer surfaces (front and back) are composed of silicon. This is done by depositing a by a polysilicon "protect" layer on the wafer. Then windows are cut into this layer where epitaxial growth is required. The polysilicon layer also serves to protect the field isolation and/or critical oxides from HF attack during the HF-passivation dip, and to reduce the overall moisture load in the UHV/CVD chamber. As a result, the basic approach is to provide a substrate which resembles as closely as possible the ideal of an all-silicon, hydrogen-passivated surface of a bare Si-wafer.

10.8 THERMAL STABILITY PROCESS INTEGRATION ISSUES

It was noted previously that it must be possible to maintain a nearly defect-free Si:Ge-strained-layer on Si throughout the entire BiCMOS process. Although unstrained Si:Ge layers also exhibit significant bandgap-reduction, the presence of an extensive misfit dislocation network in the material would preclude the fabrication of high quality bipolar devices. However, by adopting effective epitaxial growth techniques and optimizing the Si:Ge-strained-layer characteristics, it is possible to minimize strain-induced-defects. It turns out that Si:Ge films with doping levels of Ge up to 13% are able to maintain their

mechanical stability under annealing conditions of up to 900°C for 30 min, making it feasible to use such Si:Ge layers in BiCMOS processes (provided the layers are not subjected to temperatures and times above this limit).

10.9 COMPATIBILITY WITH ION-IMPLANTATION PROCESSES

There have been some questions concerning the compatibility of a Si:Ge layer with subsequent ion-implantation steps. That is, one might wonder if introducing significant amounts of lattice damage to a strained layer through implantation would degrade its integrity. Such implantation damage could provide a source of defect nucleation sites, which in turn could lead to deleterious formation of an extensive misfit dislocation network upon annealing. It turns out, however, that Si:Ge layers which are mechanically stable are apparently able to undergo implantation and anneal cycles with no additional defect-generation or movement beyond that expected in single-crystal silicon.

The two processes likely to cause problems are the moderate-energy, low-dose implants used to selectively implant dopants for the sub-collector of the HBT (phosphorus, at 100-keV, $1x10^{13}/cm^2$), and the high-dose, low-energy implants used to form the extrinsic base of the HBT and the source/drain regions of the PMOSFETs (boron, 5-keV, $1x10^{15}/cm^2$). The effects of these implants on the Si:Ge-layers was examined, and it was found that no significant yield impact was observed. A examination of the Si:Ge-strained-layer after ion implantation was studied using transmission electron microscopy (TEM). There was no evidence of misfit-dislocation formation within the device regions.

10.10 COMPATIBILITY WITH THERMAL OXIDATION PROCESSES

It must also be possible to oxidize exposed regions of Si to create an insulator in the integrated BiCMOS process. Studies indicate that the presence of Ge in the Si dramatically enhances the oxidation rate, and that Ge segregates to the growth-interface during oxidation. Since local-oxidation is generally used to provide electrical junction isolation, the Ge "pile-up" effect is likely to lead to high levels of leakage and potential reliability concerns under electrical stress. A generally accepted way to control both of these effects is to use an all-Si capping-layer for oxide growth and isolation purposes.

A low-thermal-budget CMOS process is required for thermal-cycle compatibility with a narrow-base *npn* process. After the base-epitaxial-layer is grown, the thermal-cycle should be kept to a minimum, ideally consisting of only the bipolar *npn*-emitter-cycle, which entails a furnace anneal and an RTA anneal. The two largest post-gate-etch thermal-cycles in a CMOS process are the

polysilicon reoxidation step, and the activation anneal performed after the source/drain implantation step. Both of these steps in conventional CMOS are typically furnace anneals at temperatures > 900°C, which causes significant diffusion of the base. However, the poly reoxidation step has been recently carried out at somewhat lower temperatures (850°C), and short RTP anneals have generally been replacing furnace S/D anneal steps. As the Si:Ge processes have been evolving through several generations, the thermal budget issues have been handled differently. We will discuss this subject again in Sects. 10.11 and 10.12.

10.11 VERTICAL SCALING OF Si:Ge HBTS

Using Si:Ge and bandgap-engineering, bipolar transistors will continue to provide circuit designers with the performance necessary to meet the stringent circuit and systems requirements of the next decade. In demonstration of this trend, companies are leveraging the learning from 50-GHz-f_T Si:Ge-devices to develop 75 to 90-GHz-f_T Si:Ge-devices. Continued progress in process and device design will certainly push performance of the Si:Ge-HBT to higher levels in the future.

Higher f_T performance is principally achieved through the vertical scaling of dopants and dimensions in the HBT device. Such scaling primarily involves: **1**) a reduction in transit time through the base and collector by reduction of the vertical-layer dimensions; and **2**) a commensurate increase in Ge-grading across the (now thinner) neutral base.

Key to vertical dimension reduction is the minimization of thermal cycles in post-base-deposition processing. The modular integration trend exemplified by the base-after-gate process sequence (described in Sect. 10.12) provides one important aspect to obtaining vertically scaled HBTs. In addition, movement towards RTP processing is favorable to achieving low overall thermal budgets.

Another technique for reducing the base-width-dimension is to incorporate carbon into the base epitaxy. Carbon incorporation at concentrations lower than $10^{20}/cm^3$ has been shown to significantly reduce B diffusion (by lowering the Si self-interstitial concentration within the C-rich regions).[37]

The choice of emitter dopant is closely linked to the device thermal-cycles. Various emitter dopants have been described in the literature, including implanted arsenic or phosphorus emitters, and *in situ*-doped phosphorus emitters. Because thermal-cycles required to diffuse and activate arsenic emitters will adversely impact base-widths as thinner bases are used, *in-situ*-doped phosphorus emitters will become prevalent in future Si:Ge-HBT device designs targeting greater than 100-GHz-f_T.[38]

10.12 BASE-DURING-GATE AND BASE-AFTER-GATE BiCMOS PROCESS FLOWS

As noted earlier, the goal of producing a BiCMOS process that integrates HBTs into a CMOS process flow is to ensure that the electrical characteristics of the CMOS devices remains as close to those obtained in the CMOS process alone. Therefore, the process must maintain the thermal excursions needed to produce the CMOS structures. In the first two generations of BiCMOS technologies developed at IBM this was accomplished by sharing layers and thermalcycles. That is, the HBT was fabricated *during* the fabrication of the CMOS layers, and this sequence was termed *base-during-gate* (BDGate). A flow diagram of the base-during-gate process is shown in Fig. 10-8.[29] To avoid significant widening of the base during the post-base-deposition thermal-processes of poly-reoxidation and NMOSFET-S/D-annealing, a low-temperature poly-reox-step was used, together with a low-thermal-budget CMOS-process technology. The BDGate flow worked well down to the 0.5-μm-BiCMOS technology node, but it was difficult to extend it into the next, scaled generations. This was because the thermal-cycles needed in the next generations became too high for the HBT-device.

As a result, a new flow, termed *base-after-gate* (BAGate) was developed. In the base-after-gate sequence the thermal-cycles for the CMOS-devices and the HBTs were decoupled. This is achieved by building the HBT after the CMOS is fabricated. That is, the CMOS is largely fabricated in one complete block before the Si:Ge-base is deposited. Therefore, the CMOS-technology steps can be completely copied without modification. The bipolar-transistor is fabricated in a complete block after the CMOS without sharing any thermal-cycles (except for a final emitter drive-in RTA). Since the HBT-processing is performed at low temperatures, thermal effects on the FETs are minimal. A flow diagram of the base-after-gate process for the 0.25-μm-BiCMOS process is given in Fig. 10-9. By decoupling the fabrication of the CMOS- and HBT-structures, the migration from generation-to-generation becomes easier.[30,35,38] The modularity of the BAGate process supports rapidly developing derivatives and low-cost alternative BiCMOS technologies. The baseline-CMOS processes are also incorporating more analog elements (capacitors, inductors, and resistors) because of the expanding use of CMOS for analog products. This trend is favorable for BiCMOS technology because fewer processing steps need to be added to the CMOS to have an analog-Si:Ge-BiCMOS process.

Both processes employ a planarized deep-trench-plus-shallow-trench isolation-structure fabricated with the aid of CMP technology. The Si:Ge is blanket deposited so that it is single-crystal over the exposed silicon-regions and polysilicon over the field-oxide isolation-regions. Standard CMOS processing is

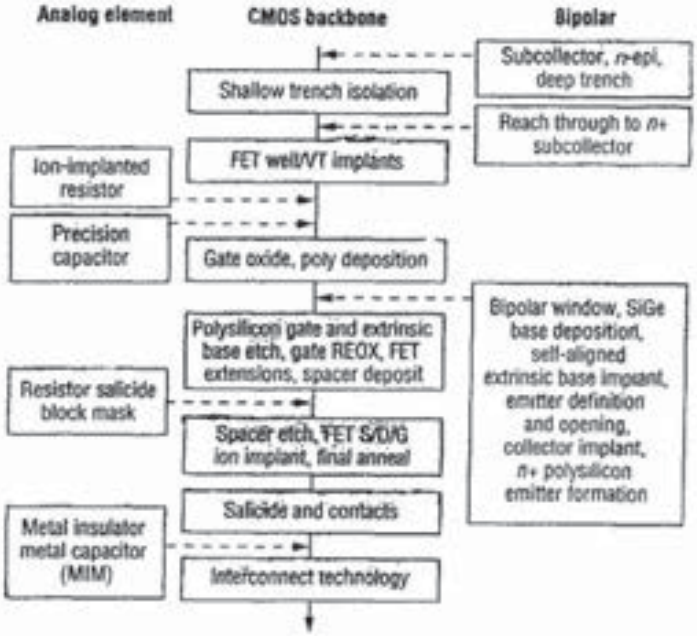


Fig. 10-8 CMOS process flow for fabricating *Base-During-Gate* Si:Ge BiCMOS structure.[29]

used, including the use of titanium salicide. The metallization is planarized Al-Cu metal, patterned by RIE, with tungsten interlevel plugs. Fig. 10-10 shows a schematic cross-section of this BiCMOS structure.

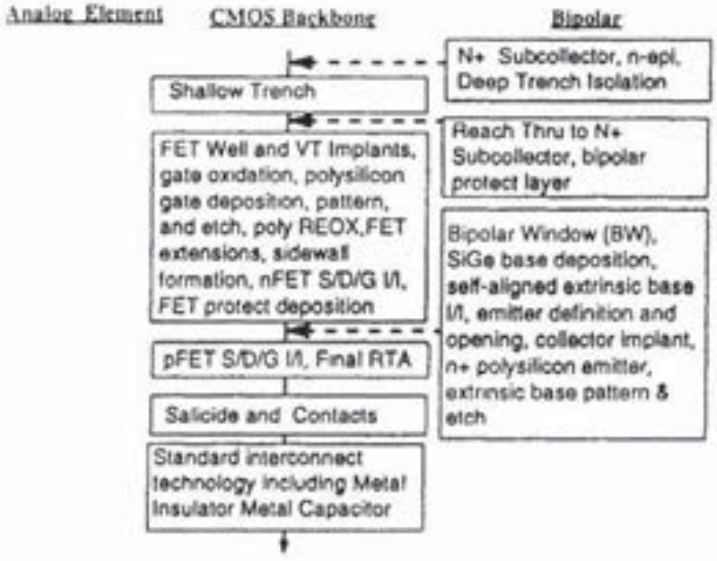


Fig. 10-9 CMOS process flow for fabricating *base-after-gate* Si:Ge BiCMOS structures.[29]

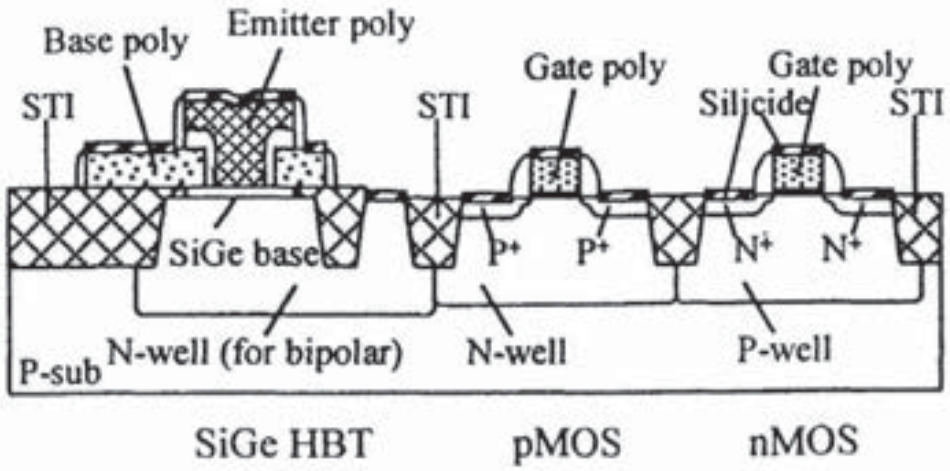
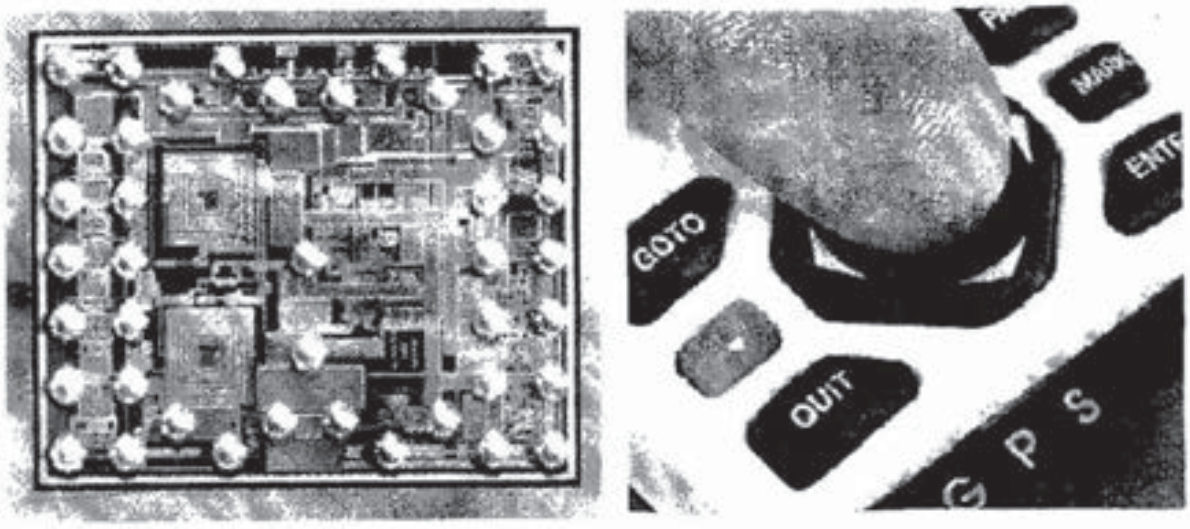


Fig. 10-10 Schematic cross-sectional view of the Si:Ge BiCMOS. The Si:Ge HBT has a self-aligned structure.[16]

10.13 EXAMPLES OF THE MODULAR INTEGRATION Si:Ge-HBTS INTO A BiCMOS PROCESS FLOW

The first Si:Ge-IC that IBM offered commercially was built using 0.25-μm geometries and combined 1200 Si:Ge-bipolar-transistors with over million CMOS-transistors. This IC was the highest-performance, lowest-noise *data-channel-circuit* in the world. It is used to read data from high-speed hard-drives. The IBM fab at Burlington VT was processing 75 wafers per day of this device (starting in 1999), with yields above 75%.[20]



a) b)

Fig. 10-11 a) Single-chip GPS Receiver; b) Relative size of GPS Receiver.[8]

In 1998 IBM described the characteristics of a 1.8-million-transistor ASIC chip, fabricated using a Si:Ge-BiCMOS technology.[31] Another description of the modular integration of an HBT that used a Si:Ge-base-layer (in which carbon was also incorporated) into a standard 0.25-μm-CMOS process to yield a BiCMOS technology is given in Ref. 32. Neither the HBT nor the CMOS parameters were adversely impacted by the additional thermal budget of the complete BiCMOS process. The technology is low-cost, has high-performance, and allows re-use of existing CMOS designs. The Si/Si:Ge-HBTs in this technology are able to perform at very high speeds; f_{max} = 65-GHz. Because such high device-switching-speeds are not necessary for most wireless circuits operating at 900-MHz and 2.4-GHz, the Si:Ge leverage comes from being able to trade the excess speed for improvement in other device attributes; most notably, low-power operation (Fig. 10-11). Another example of integrating Si:Ge-HBTs into a CMOS process is given in Ref. 33. In this process four additional masking steps were needed to create this BiCMOS process. The result is a BiCMOS technology that includes a super self-aligned (SSA) Si:Ge-HBT with a peak f_T of 52-GHz and a peak f_{max} of 70-GHz. A 0.18-μm rf-BiCMOS technology with a 73-GHz f_T S i:Ge-HBT was described in late 2000.[18] A two-step annealing technique is used to solve the thermal budget tradeoff between Si:Ge HBTs and CMOS. IBM has also devoted significant effort to develop high-quality capacitors and inductors, which can be integrated into the Si:Ge-process. These passive components allow IBM to build the world's only single chip Global Positioning Satellite (GPS) receiver (Fig. 10-11).[8]

A recent report described the effects of scaling on the performance characteristic of deep-submicron Si:Ge HBTs (Fig. 10-12).[41]

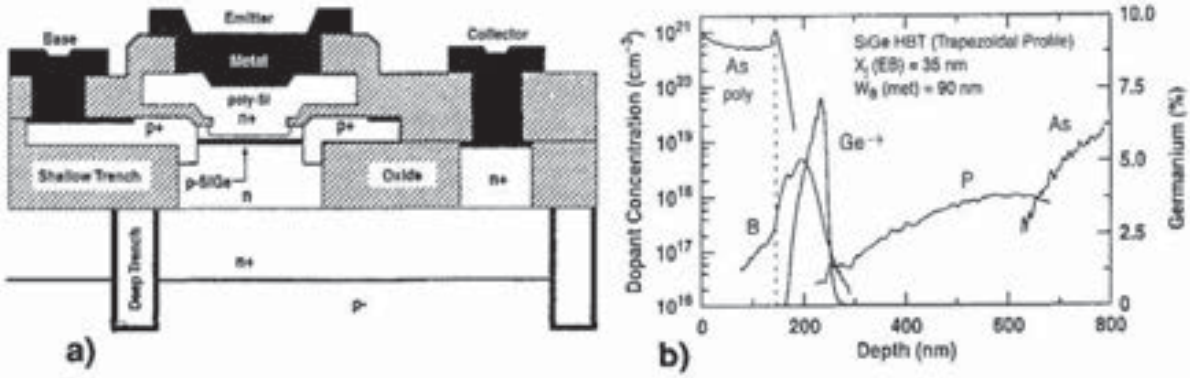


Fig. 10-12 a) Schematic cross-section of deep-submicron Si:Ge HBT; b) SIMS doping profiles of the dee-submicron Si:Ge HBT.[41] (© 2002 IEEE)

REFERENCES

1. M. Helms, Chip: Semiconductor Industry Source Book, 1999, pp. 26-27, www.bps-IT.com.
2. D.C. Ahlgren *et al.*, "Manufacturability of an Integrated SiGe HBT Technology for the Analog and Wireless Marketplace," *Tech Dig. IEDM* 1996, p. 859.
3. S. St. Onge *et al.*, "A 0.24-μm SiGe BiCMOS Mixed Signal RF Technology Featuring a 47 GHz f_T HBT and 0.18 μm-L_{eff} CMOS," BCTM Proc. 1999.
4. G. Niu, *et al.*, "SiGe Profile Design Tradeoffs for RF Circuit applications," *Tech. Dig. IEDM* 1999 p. 573.
5. K. Washio *et al.*, "A 0.2 μm Self-Aligned SiGe HBT Featuring 107-GHz f_{max} and 6.7-ps ECL," *Tech. Dig. IEDM* 1999 p. 558.
6. A. Catalano, *Solid State Technology,* September 2000, p. 59.
7. Y.-C. Chao *et al.*, "Enhanced Performance of Sub-100 nm CMOSFETs Using Strained Epitaxial Si-Ge," *Tech. Dig. IEDM*, 2000 p. 753.
8. L.E. Larson, "Device and Technology Requirements for Next Generation Communications Systems," *Tech. Dig. IEDM*, 2000 p. 737.
9. J. Woo and J. Plummer, *IEEE Trans. Electron Dev.,* August 1988, p. 1311.
10. S.L Wright *et al.*, "Molecular Beam Epitaxy Growth of GaP on Si," *J. Appl. Physics*, Vol. 55, April 1984, p. 2916.
11. Y.H. Kwark and R.M. Swanson, "*n*-Type SIPOS and Poly-Silicon Emitters," *Solid-State Electronics,* Vol. 30, November 1987, p. 1121.
12. M. Takahashi *et al., IEEE Electron Dev. Letts.*, October 1987, p. 475.
13. T. Sugii *et al., IEEE Electron Dev. Letts.*, February 1988, p. 87.
14. H. Fujioka *et al., tech. Dig. IEDM*, 1988, p. 574.
15. D. L. Harame *et al.*, "Si/Si-Ge Epitaxial-Base Transistors - Part I: Materials, Physics and Circuits," *IEEE Trans. Electron Dev,*. Vol. 42 March 1995, p. 455.
16. B.S. Meyerson, "Low-Temperature Silicon Epitaxy by Ultrahigh Vacuum Chemical Vapor Deposition," *Appl. Phys. Letts.*, Vol. 47, p. 797 (1985).
17. J.F. Gibbons *et al.*, "Limited Reaction Processing: Silicon Epitaxy," *Appl. Phys. Letts.*, Vol. 10, p. 721 (1985).
18. T. Hashimoto *et al.*, "A 73 GHz 0.18 μm f_T RF-Si:Ge BiCMOS Technology Considering Thermal Budget Trade-Off and With Reduced Boron-Spike Effect on HBT Characteristics," *Tech Dig. IEDM*, 2000, p. 149.
19. R. People and J.C. Bean, *Appl. Phys. Letts.*, **48**, p. 538 (1986).
20. B.S. Meyerson, "Silicon Germanium Technology: A 'Material of the Future' Gone Mainstream," *IEEE/SEMI Advanced Semiconductor Manufacturing Conference,* Fall 2000, Boston MA.
21. B.S. Meyerson, "Low-Temperature Si & SiGe Epitaxy by Ultrahigh Vacuum/CVD: Process Fundamentals," *IBM J. Res. Dev.,* Vol. 34, p. 806 (1990).

22. K.E. Violette *et al*, "Low-Temperature Selective Silicon Epitaxy by Ultrahigh Vacuum Rapid Thermal Chemical Vapor Deposition Using Si_2H_6, H_2 and HCl," *Appl. Phys. Letts.,* Vol. 68, p. 66 (1996).
23. "Motorola's Next Si:Ge Wireless Technology Will Use the ASM Epsilon E2000 Reactor." *Press Release: ASMl*, March 1, 2001, Bilthoven, Netherlands.
24. D.J. Meyer and J. Italiano, "Productive Single-Wafer Si:Ge Processing," *Meeting Abstracts of the Fall Mtg. 1999, Electrochem. Society*, Abs. No. 1314.
25. ibid. Ref. 6.
26. J. Bock *et al.,* "Si-Ge Bipolar Technology for Mixed Digital and Analogue RF Applications," *Tech. Dig. IEDM* 2000 p. 745.
27. K. Washio *et al.,* "A 0.2 μm 180-GHz - 6.7 ps-ECL SOI/HRS Self-Aligned SEG Si-Ge HBT/CMOS Technology for Microwave and High-Speed Digital Applications," *Tech. Dig. IEDM* 2000 p. 741.
28. D. L. Harame *et al.,* "Si/Si-Ge Epitaxial-Base Transistors - Part II: Process Integration and Analog Applications," *IEEE Trans. Electron Dev.* **42** Mar 1995, p. 469.
29. D.C. Ahlgren and B. Jaganathan, "SiGe for Mainstream Semiconductor Manufacturing," *Solid State Technology,* January 2000, p. 53.
30. D. Harame *et al.,* "Trends in SiGe BiCMOS Process Integration," *Meeting Abstracts of the Fall 1999 Mtg. Electrochemical Society,* Abs. No. 1333.
31. R.A. Johnson *et al.*, "1.8 Million Transistor CMOS ASIC Fabricated in a Si-Ge BiCMOS Technology," *Tech. Dig. IEDM* 1998, p. 217.
32. K.E. Ewald *et al.,* "Modular Integration of High-Performance SiGe:C HBTs in a Deep Submicron, Epi-Free CMOS Process," *Tech Dig. IEDM* 1999 p. 561.
33. C.A. King, "Very-Low Cost Graded SiGe-Base Bipolar Transistors for a High Performance Modular BiCMOS Process," *Tech Dig. IEDM* 1999 p. 565.
34. S.R. Stiffler *et al.,* "The Thermal Stability of Si-Ge Films Deposited by UHV/CVD," *Appl. Phys. Lett.*, **70**, p. 1416, 1991.
35. D.C. Ahlgren *et al.*, "Si-Ge Fuels High-Speed Communications," *Semiconductor International,* November 2000, p. 62.
36. D.L. Harame and. B.S, Meyerson, "The Early History of IBM's SiGe Mixed Signal Technology," *IEEE Trans. Electron Dev.,* **48**, November 2001, p. 2555.
37. H. Rucker *et al.,* "Dopant Diffusion in C-doped Si and SiGe: Physical Model and Experimental Verification," *Tech. Dig. IEDM*, 1999, p. 345.
38. D.L. Harame *et al.,* "Current Status and Future Trends in SiGe BiCMOS Technology," *IEEE Trans. Electron Dev.,* **48**, November 2001, p. 2575.
39. A.V. Samilov *et al.,* "Dedicated Pre-Clean Reduces Epi's Thermal Budget," *Semiconductor International,* November 2000, p. 73.
40. V. Loup *et al.,* "Fine Control of Low-Temperature Epitaxial Growth," *Solid State Technology,* July 2001, p. 91.
41. S.Zhang *et al.,* "Effects of Geometrical Scaling on Frequency Response and Noise Performance of SiGe HBTs," *IEEE Trans. Electron Dev.* Mar. 2001, p. 429.

Chapter 11

SILICON-ON-INSULATOR (SOI) TECHNOLOGY

In this chapter the characteristics of silicon-on-insulator (SOI) technologies fabricated with deep-submicron MOSFETs are described. The focus is on the SOI-devices themselves as well as on the process technologies used to fabricate both the starting SOI-wafers and the SOI ICs. An argument is also presented as to why SOI technology may provide a path for continuing to advance CMOS ICs for several more generations. That is, implementation of SOI may make it possible to exceed the limits facing bulk-silicon-CMOS in the near term.

11.1 WHAT IS SILICON-ON-INSULATOR (SOI)?

Silicon-on-insulator (*SOI*) circuits consist of ICs made from single-device islands dielectrically isolated from the substrate and each other. Each such island contains a single transistor, with the isolated transistor being either a MOSFET or a bipolar transistor. However, since about 90% of the ICs manufactured in 2002 are based on CMOS-technology, the discussion in this chapter will be restricted to SOI technology involving MOSFETs. Figure 11-1 shows a cross-section of a MOSFET fabricated on an SOI wafer that also uses STI as the isolation structure to provide lateral isolation. Note that in the case of a MOSFET fabricated on an

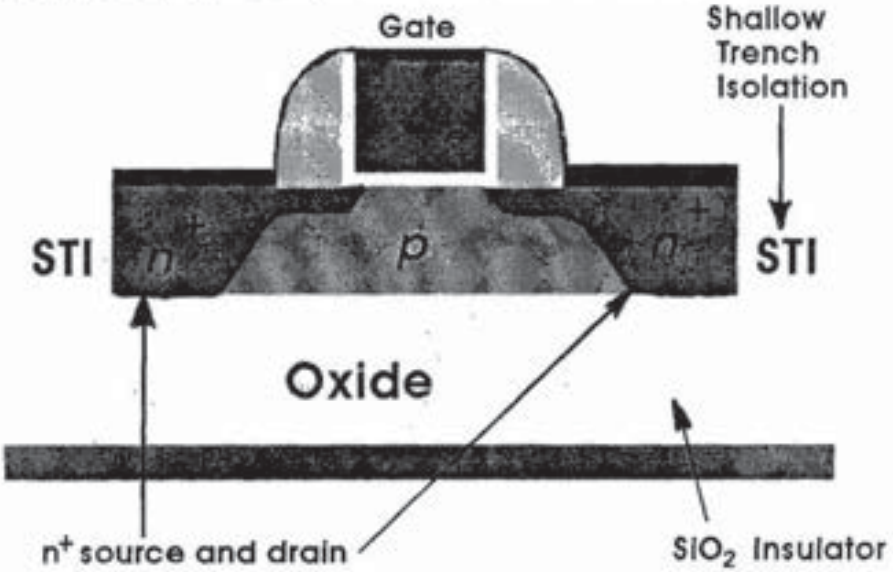


Fig. 11-1 Cross-section of a MOSFET fabricated on an SOI wafer.

SOI wafer, a *buried oxide* (*BOX*) underneath the device provides vertical isolation, and a trench oxide (in this example), provides the lateral isolation. Since the trench oxide extends all the way through the silicon device layer atop the BOX (and a thermal oxide exists on the wafer top surface), dielectric material continuously surrounds each MOSFET on all sides. Thus, complete electrical isolation is provided for the MOSFET with respect to all other devices, and the substrate, on the wafer.

11.1.1 Isolation Issues in CMOS Fabricated in Bulk-Silicon Wafers

The electrical isolation requirements of CMOS ICs built in SOI are quite different from those encountered when CMOS ICs are fabricated in bulk-silicon. To see why, the circuit strategies and structures needed for isolation purposes in CMOS circuits on bulk-silicon must be examined. For reference purposes a single-MOSFET built in a bulk-silicon-substrate is shown in Fig. 11-2a (with various device structures and regions defined). Figure 11-2b shows two NMOSFETs built on a *p*-doped substrate, together with the parasitic-transistor that exists in the field-region separating the active devices. Semirecessed-LOCOS is used to produce the field-oxide (which is part of the device structure employed to maintain electrical isolation among such MOSFETs). Figure 11-2c shows an *n*-well CMOS structure fabricated in a *p*-type bulk-silicon wafer. This figure also shows the parasitic MOSFETs that are associated with that structure (as well as the *pn*-junction that separates the *n*-wells from the *p*-substrate). Again, a semirecessed-LOCOS field-oxide plays a role in maintaining electrical isolation among the MOSFETs of the CMOS structure. (It should be noted that shallow-trench isolation [STI] is used in place of LOCOS to form the field-oxide in CMOS technologies below 0.35-μm. Nevertheless, the isolation issues described here for CMOS on bulk-silicon remain the same, regardless how the field-oxide is formed). Finally, in Fig. 11-2d a cross-section of a CMOS IC structure is depicted, together with a schematic depiction of the parasitic-bipolar-transistors and associated lateral currents.

Using the IC device structures described above, the isolation issues involved in CMOS ICs can be undertaken. The first item to be noted is that the *pn*-junction separating a well from the substrate must be kept under reverse-bias to prevent current from flowing from the power-supply-line to ground through forward-bias of this junction. Second, inter-device isolation for devices within a well (*intra-well isolation)* or within the substrate regions, entails the use of a field-oxide to suppress spurious surface-leakage effects between active devices within a well. Third, a field-oxide and channel-stop implants are also used to prevent leakage

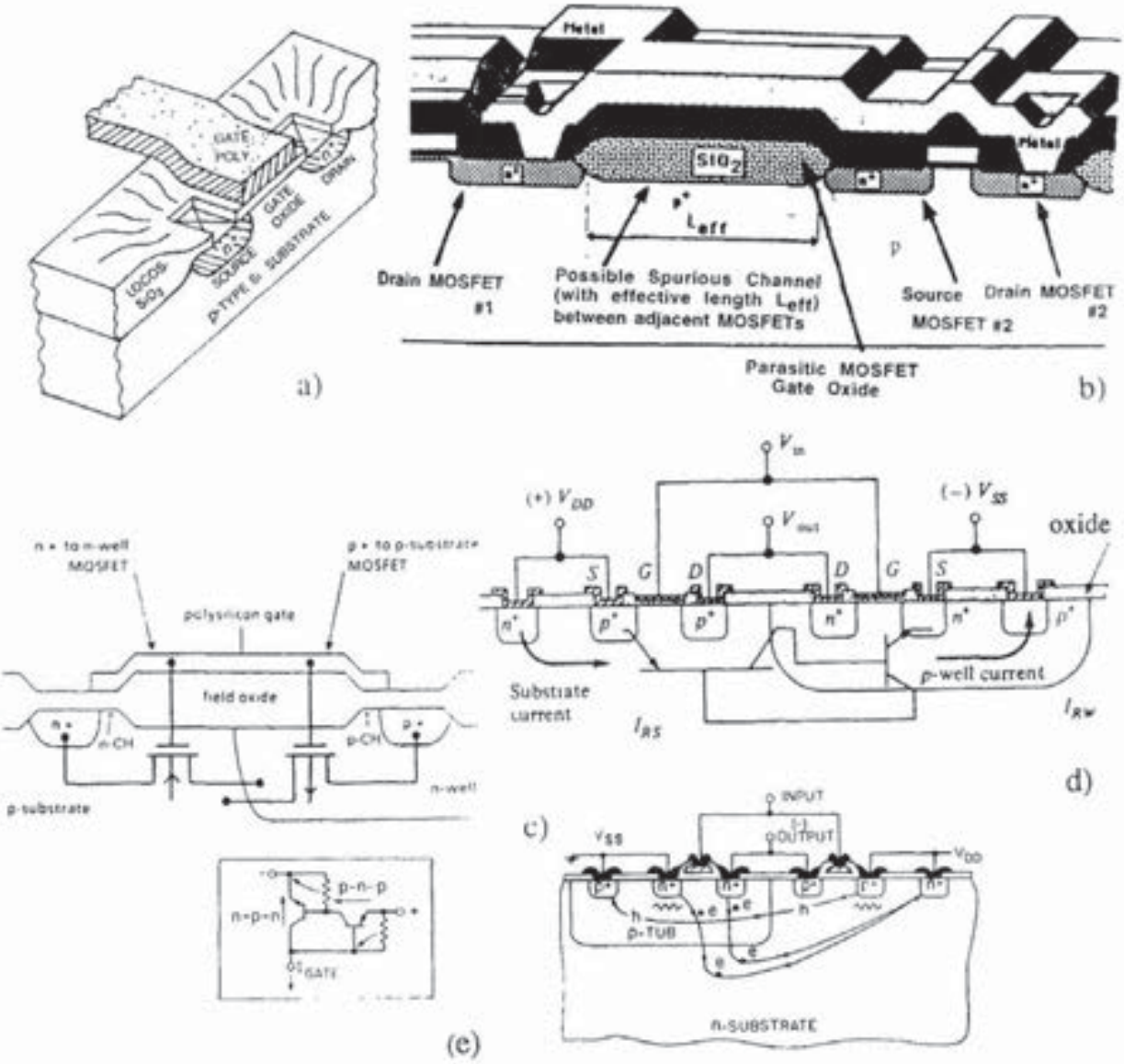


Fig. 11-2 a) A single MOSFET built in a bulk-silicon substrate, with various device structures and regions defined; b) Two MOSFETs (e.g., both NMOSFETs) built in a substrate having one type of doping (i.e., on a *p*-doped substrate), together with the parasitic transistor that exists in the field-region separating these MOSFETs. Semirecessed-LOCOS is used to produce the field-oxide (which is part of the device structure employed to maintain electrical isolation among such MOSFETs); c) An *n*-well CMOS structure fabricated in a *p*-type bulk-silicon wafer. Also shown in the figure are the parasitic MOSFETs that are associated with that structure, as well as the *pn* junction the separates the *n*-wells from the *p*-substrate; d) A cross-section of a CMOS-IC-structure is depicted, together with a schematic depiction of the parasitic-bipolar-transistors and associated lateral currents; e) The parasitic bipolar transistors that exist in CMOS structures can be *turned on*, which can cause the CMOS IC to go into a latchup state.

between devices among different wells (*inter-well isolation*). Yet, despite employing these isolation structures, intra-well- and inter-well-leakage still occurs to some degree in CMOS ICs made in bulk-silicon (and the problem gets worse as the devices are brought closer together). Note that Volume 3 can be consulted for a more detailed discussion of isolation issues in CMOS built in bulk-silicon wafers. In addition, the parasitic-bipolar-transistors that exist in CMOS structures can be *turned-on*, and this can cause the CMOS IC to go into a *latchup state* (Fig. 11-2e). That is, latchup can occur because two parasitic-bipolar-transistors are always simultaneously formed in the course of fabricating a CMOS structure. In bulk-CMOS, these two transistors are connected in such a way that if the current gain of the loop formed by these two devices is larger than unity, latchup can occur. Latchup can be triggered by different mechanisms, such as node- voltage overshoots, displacement-current, junction-avalanching, and pho-tocurrents. It is true that latchup is a well-understood phenomenon which can be dealt with in bulk-CMOS using proper processing and design methodologies. However, with the constant reduction in feature spacing, latchup immunity in bulk CMOS ICs remains a continuing challenge.

11.1.2 Isolation in SOI CMOS

The isolation issues of CMOS fabricated in SOI wafers are much less complex than those involving CMOS ICs in bulk-silicon. This is because each MOSFET in SOI is built in an island of silicon on an insulating substrate. Hence, all of the previously-mentioned possible isolation causes are eliminated. Let us explain this assertion in more detail.

First, the requirement of keeping the *pn*-junction that separates the CMOS well regions from the substrate under reverse-bias is eliminated. That is, this requirement no longer exists because a continuous oxide barrier exists in SOI between the well-regions and the substrate. Consequently, current cannot flow between the power and ground lines, even if the junction in question becomes forward-biased. Second, the possibility of leakage current between devices within a well (intra-well-leakage) and from well-to-well (inter-well-leakage) is eliminated. This is because the parasitic-field-region-MOSFET does not exist (as a result of the field oxide extending through the silicon film and joining the buried oxide). In other words, the channel between the source and drain of any parasitic-field-region-MOSFET is electrically interrupted by the oxide moat surrounding each MOSFET within an SOI IC. Consequently, leakage current due to: 1) field-inversion of the channel of a parasitic-field-region-MOSFET; or

Table 11-1 ITRS Roadmap of the Required MOSFET Junction Characteristics

Year	2000	2001	2002	2003	2004	2005	2006	2010	2014
Gate size (nm)	120	100	85	80	70	65	45	32	22
S/D junction depths (nm)									
Contact	65–125	55–105	45–90	43–8	38–75	35–70	30–55	20–40	15–35
Extension/channel	36–60	30–50	25–43	24–40	20–35	20–33	16–26	11–19	6–13
SSRD channel	18–30	15–25	12–22	12–20	10–18	10–17	8–13	5–10	4–8
Extension doping									
Sheet resistance (Ω/sq.)	310–760	280–730	250–700	240–675	220–650	200–625	150–525	120–450	100–400
Resistivity (10^{-3}Ω-cm)	1.1–4.6	0.84–3.6	0.62–3.0	0.57–2.7	0.44–2.3	0.40–2.1	0.24–1.4	0.13–0.86	0.08–0.52
Lateral extension junction abruptness (nm/decade)									
For V_t-roll-off control	12	10	8.5	8	7	6.5	4.5	3.2	2.2
For spreading resistance control	4.6	4.1	3.4	2.9	2.7	2.4	1.25	0.8	0.5

2) parasitic-MOSFET subsurface-punchthrough, do not exist in SOI-technologies. Finally, the existence of vertical- (as well as lateral-) isolation within SOI ICs eliminates the possibility of latchup. That is, the current paths from the collectors of the parasitic-bipolar-transistors to the bases of the complimentary bipolar-transistors (which set up feedback paths in CMOS in bulk-silicon that cause the circuits to enter latchup), are interrupted by the vertical oxide of the SOI. Hence, latchup can no longer occur in SOI ICs.

Despite the fact that intra-well and inter-well leakage paths are interrupted in SOI CMOS structures, new *intra-device* parasitic transistor paths inherent in SOI CMOS structures can still produce leakage current, as is discussed in Sect. 11.7.

11.2 THE "END OF THE ROAD FOR BULK CMOS" SCENARIO PREDICTED BY THE 2001 INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

The *2001 International Technology Roadmap for Semiconductors* (ITRS-2001) indicates that planar CMOS transistors on bulk-silicon are expected to reach their limits at gate lengths of about 50-nm in the 2005-2006 time frame. That is, the ITRS-2001 analysis documents a clear "end-of-roadmap" event-horizon at the onset of IC device scaling to 50-nm gate-lengths.[1] The premise of the ITRS study was to predict, as far as possible, the characteristics of an extension of the industry's "standard technology," namely, planar-CMOS-transistors fabricated

on bulk-silicon wafers. The study concludes that many process and materials constraints combine to force this change in the technology path. That is, the conclusions drawn in this study are that sometime between 2005 and 2008 nearly all of the 30 or so transistor characteristics required for the extension of Moore's Law are listed as "no known solutions." This is a clear signal that IC transistor technology requires fundamental changes in the next half-decade. In this section the principal issues limiting *junction-formation* in sub-100-nm CMOS ICs on bulk-silicon will be detailed. As part of the discussion to be presented in the next section (Section 11.3), it will be described how some of these constraints can be removed or relaxed through the use of SOI. That is, by implementing CMOS-on-SOI a path may be provided to allow CMOS ICs to be fabricated beyond the 50-nm gate-length-limit faced by CMOS-on-bulk-silicon. An extension of as much as a decade beyond the "end-of-the roa-map" scenario posited by the ITRS-2001 may be possible by fabricating CMOS ICs on SOI-wafers.

The three principal challenges for forming shallow-junctions in CMOS on bulk-silicon wafers are as follows:

1. Limiting dopant-diffusion in the source/drain regions to shallow depths;
2. Activating a sufficient percentage of the dopants implanted into the source/drain regions to provide source/drain regions that exhibit sufficiently low contact-resistance and sheet-resistivities:
3. Obtaining source/drain junctions that have highly-abrupt vertical and lateral doping-profiles.

The ITRS-2001 values for these key transistor parameters, as a function of the device generation (which correlates to the gate-length of that generation and the expected time of its introduction) are given in Table 11-1.

11.2.1 Source/Drain Junction Scaling

As MOSFET gate lengths (L_L) are scaled, the other device structures (i.e., the gate-oxide-thickness and the junction-depth of the source/drain extension) are also scaled proportionally. To maintain such "proportional scaling" for CMOS when the gate-length decreases below 100-nm, the design rule used in the ITRS-2001 study was that the junction depth of the source/drain extension $X_{j(extension)}$ would scale according to: $X_{j(extension)} \sim 0.55 \times L_L$ within a range of ±25% (see Fig. 11-2a for a graphic that depicts the region referred to as the source/drain extension). Controlling junction depths to less than 55-nm for sub-100-nm MOSFETs will push the limits of sub-keV boron implantation and "spike" anneals.[2]

11.2.2 Dopant Activation in Shallow Source/Drain Junctions

In the ITRS-2001 model, it is expected that ways will be found to produce highly doped source/drain extension-regions having a sufficiently low resistivity (to reduce the sheet-resistance of these regions) at the same time the depth of the junctions of these regions is reduced. If this can be done (i.e., if the resistivity of the source/drain extension regions could somehow be reduced from the value of 10^{-3} Ω-cm [that is produced in the source/drain extension-regions of 0.15-μm NMOS transistors by using low-energy boron implantation {0.2-2.0-keV} and "spike" anneals] to say 10^{-4} Ω-cm) a sheet-resistance versus junction-depth curve (as plotted in Fig. 11-3 for the value of $\rho = 10^{-4}$ Ω-cm) could be obtained.[3] Overlaid on these sheet-resistance versus junction-depth curves (with ρ as the parameter), are the industry expectations of the value of ρ for: **1**) high-performance ICs; and **2**) cost-focus (i.e., lower-cost) ICs, as taken from the ITRS-2001. That is, the values of ρ that the ITRS-2001 roadmap expects in order to be able to fabricate CMOS ICs on bulk-silicon for smaller gate-length-devices, decreases from about 10^{-3} Ω-cm for NMOSFETs with gate lengths of 0.14-μm, down to

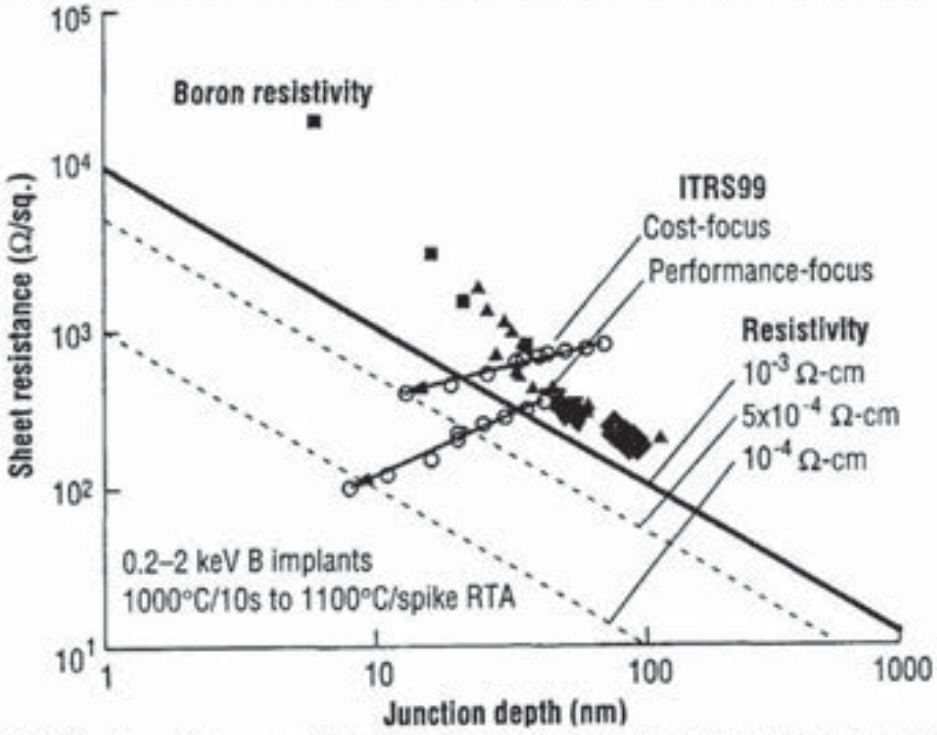


Fig. 11-3 Sheet-resistance and junction-depths for low-energy boron implants and spike-implant anneals, and industry expectations for S/D extension junctions (starting with 140-nm gates to 22-nm gates at the head of the arrows) for cost-focus and performance focus devices. The dopant resistivity for sub-keV boron implant and spike-RTP process (in CZ-Si) is limited by a "wall" at about 10^{-3} Ω-cm.

about 10^{-4} Ω-cm for NMOSFETs with 22-nm gate lengths (corresponding to a junction-depth of 8-nm and a sheet resistance of 100 Ω/sq). Some alternative doping technology may be required to achieve such low resistivity. Possible alternative doping techniques include: **a**) low-temperature (or selective) epitaxy; **b**) diffusion from a doped-glass source; **c**) diffusion from a low-temperature Si-Ge alloy; **d**) direct gas-source-diffusion in an RTP-system; **e**) gas-immersion-laser-diffusion (GILD); and **f**) planar-doping-layers using atomic-layer-epitaxy (ALE). Of these many possible processes, low-temperature-epitaxy may hold some key advantages, as it can also be done in a selective manner (which may possibly minimize masking steps). However, with selective epitaxy, separate hard-masks are required for *n*- and *p*-S/D-regions. On the other hand, GILD with projection masking (PGILD) does not require resists or hard-masking and may become an important doping technology.

At this time, however, it is not clear whether any of the above methods of reducing the value of ρ in the source/drain extension regions of NMOSFETs will be able to reduce ρ down to 10^{-4} Ω-cm. In fact, the resistivities of junctions formed by the present day "best methods" of sub-keV boron implants and spike anneals appears to be limited by the "activation-wall" to a value no smaller than 10^{-3} Ω-cm.[4]

11.2.3 Doping-Profile Abruptness in Shallow Source/Drain Junctions

To permit sub-100-nm MOSFETs to exhibit the performance characteristics expected in the ITRS-2001 roadmap, the total source-to-drain resistance of such MOSFETs must be kept sufficiently small.[5] Controlling this total source-to-drain resistance-value in deep-submicron MOSFETs requires that special attention be paid to the contact-resistance and spreading-resistance at the junction of the source extension and the channel. As shown in Fig. 11-4, parasitic-resistances exist in the path between the metal-to-Si substrate interface and the region in the device where the actual transistor action begins (see also Sect. 5.4, Chap. 5).

In the planar MOSFET, the device-region in which transistor-action occurs is the channel-region under the gate. The current flowing from the metal-interconnect lines to the channel-region must be transported from the metal-Si contact to the edge of the channel through the source or drain regions (Fig. 11-4). In the MOSFET, current enters the contact perpendicular to the wafer surface, and then travels parallel to the surface in order to reach the channel. The total parasitic-series-resistance, R_S, of the current path from the contact to the edge of the channel can be modeled by the sum of four components,

$$R_S = R_{co} + R_{sh} + R_{sp} + R_{ac} \qquad (11.1)$$

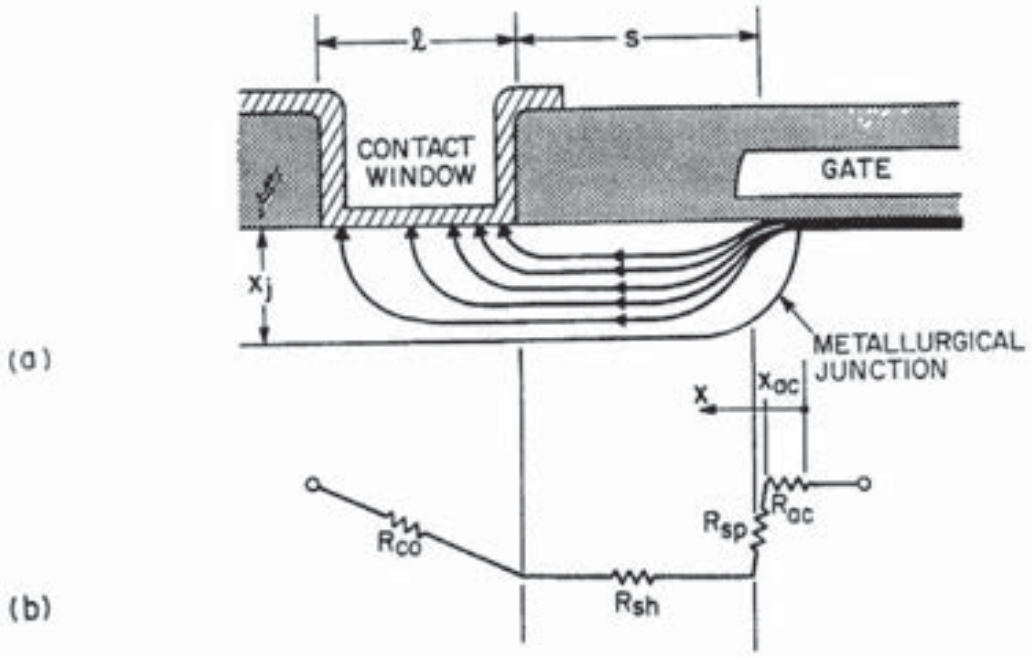


Fig. 11-4 Schematic drawing showing: a) current paths within the source and drain regions, of a planar MOSFET and; b) their corresponding representative series resistance components.[5] (© 1979 IEEE)

where R_{co} is the *contact-resistance* between the metal and the source/drain region, R_{sh} is the *sheet-resistance* of the bulk-region of the source and drain, R_{sp} describes the *resistance of the current lines crowding near the channel end of the source*, and R_{ac} is the *accumulation-layer resistance*. Qualitatively, for a non-abrupt junction, R_{sp} and R_{ac} arise because, after leaving the channel, the current does not immediately spread into the bulk of the source-drain regions (the parasitic-resistance-effect in these bulk regions is R_{sh}). Thus, R_{sp} and R_{ac} exist because the accumulation-layer conductance is much higher than the semiconductor conductance in the vicinity of the junction (see again Sect 5.3).

The key issue for sub-100-nm MOSFETs is the abruptness of the doping-profile at the lateral edge of the source extension and its effect on R_{sp} and R_{ac}. The ITRS-2001 study calls for lateral-doping-profiles that decrease into the channel region at a slope of ~3 nm/decade in concentration, for 100-nm gates. Obtaining a lateral-doping-profile abruptness of less than 3-nm/decade presents severe challenges for bulk silicon doping by implantation and rapid thermal annealing. As-implanted profiles for 0.2-keV boron into pre-amorphized silicon have a vertical profile steepness of 2.5-nm/decade.[6] The key S/D-extension profile in the lateral direction, still remains to be characterized, and will be further broadened by thermal annealing in actual transistors.

11.3 WHY SOI?

Silicon-on-insulator (SOI) technology has long been used in many specialized applications, such as radiation-hardened or high-voltage integrated circuits. It is only in recent years (since about the mid-1990s), however, that SOI has emerged as serious contender for low-power, high-performance applications.[7] The power consumption, P, in CMOS ICs is given approximately by:

$$P \sim C V^2 f + V I_{leakage} \quad (11.2)$$

where C is the total capacitance, V is the supply voltage, f is the frequency, and $I_{leakage}$ is the standby-leakage-current. Thus, in low-power applications the supply voltage must be reduced to decrease both terms of this equation. This power-supply reduction has had to follow the evolution in battery-voltage and capacity. The SEMATECH roadmap target has been to reach 0.9V by the end of the century. However, at such a low voltage, the IC performance is also reduced due to a smaller value of I_D in MOSFETs. As will be seen, however, the use of SOI allows the drain current to be significantly increased, even at such low voltages. SOI therefore also becomes appropriate for emerging system-on-chip ICs used for smart power, as well as for micro-electromechanical systems (MEMS), mixed-signal, and integrated-optics applications.

SOI also provides a path to permit the scaling of ICs to go forward. That is, to continue to increase performance and keep up with Moore's Law, several cumulative approaches have been adopted. The evolution of IC technology in the near future which uses this strategy is summarized in Fig. 11-5. This diagram indicates that integrating copper wiring and SOI will represent the next step toward the continued improvement in IC performance. Here we list the advant-

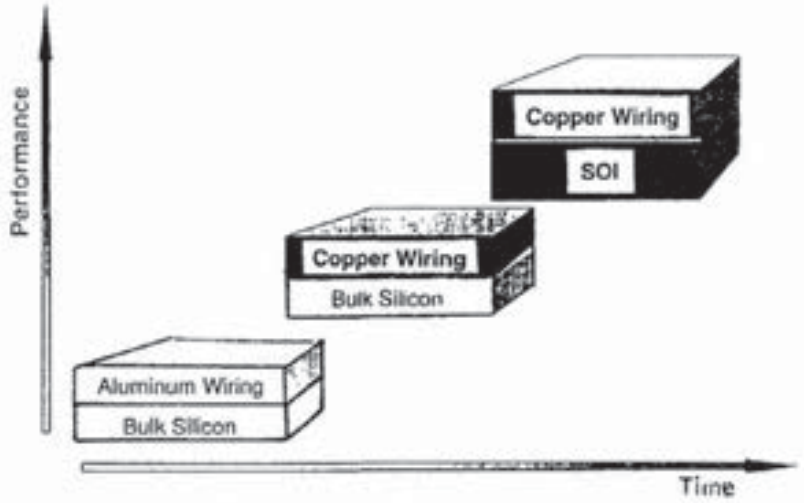


Fig. 11-5 Technology migration for future IC progress.

ages offered by CMOS-onSOI which contribute to such performance improvement.

1. **Improved Speed-Performance.** The speed performance of SOI ICs is enhanced by 20-35% compared to the same circuits fabricated in CMOS on bulk-silicon if they are operated at the same voltages. (The reasons for this increase in speed performance is due to a number of factors that are listed later as other advantages of SOI.) Thus, SOI circuits built using the design rules of "generation-n" of some CMOS technology in bulk silicon will operate as well as the ICs of "generation-n + 1" of CMOS-on-bulk-silicon. This also implies that if a roadblock occurs in some upcoming generation of CMOS on bulk silicon, by having the capability to produce these ICs in SOI, it will be possible to fabricate ICs that allow that roadblock to be exceeded for at least one generation (if not more).

2. **Reduced Power-Consumption.** If CMOS-ICs in bulk-silicon and CMOS ICs in SOI are operated to give the same speed-performance, there is a 35-70% reduction in power-consumption in the CMOS-on-SOI ICs. This advantage provides CMOS-on-SOI a huge competitive edge for low-voltage/low-power applications operated with a one-battery supply. Figure 11-6 depicts this performance-versus-power tradeoff issue.

3. **Reduced Junction-Capacitance.** One reason for the increased performance of CMOS-on-SOI is that the junction-capacitance of the MOSFETs is significantly reduced. For example, a depletion-capacitance exists at the drain-substrate junction, and its value is a function of the

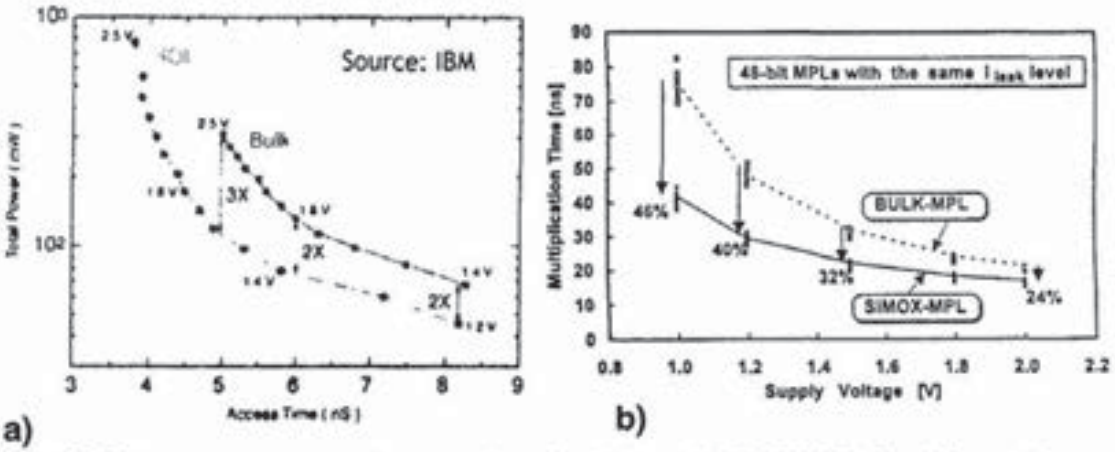


Fig. 11-6 a) Power drain and access time for bulk and SOI CMOS for drive voltages ranging from 1.2 V; b) Comparison of 48-bit-multiplier performance between bulk and SOI using a 0.25-μm CMOS technology.

substrate doping-concentration. The higher the substrate doping-concentration, the higher the capacitance. Modern submicron circuits increase the substrate doping-concentrations with each new technology generation, which contributes to an increased junction-capacitance. However, in SOI circuits, the maximum capacitance between the junctions and the substrate is the capacitance of the buried-insulator. This capacitance is proportional to the dielectric constant of the capacitance material. Silicon dioxide (which is used as the buried insulator), has a dielectric constant $k_{SiO_2} = 3.9$ which is three times smaller than that of silicon $k_{Si} = 11.7$ Therefore, a source/drain junction located on a buried-oxide gives a parasitic-capacitance approximately three times smaller than that of a bulk-junction. Since the SOI buried-insulator-thickness does not scale with device scaling, parasitic-capacitances do not necessarily increase as SOI technology progresses.

As a result of the above phenomena, and depending on the circuit layout, the total capacitance of a CMOS SOI-IC can be reduced by 15-50% compared with the same IC fabricated in bulk-silicon. In fully-depleted SOI (FD-SOI) MOSFETs (described in Sect. 11.5.1), the source/drain-regions butt-up against buried oxide (BOX), and thus no *pn*-junction exists at that device-region boundary (and hence there is no depletion-region that needs to charged and discharged there). Even in partially-depleted SOI devices (PD-SOI), the depletion-region of the junction of the source/drain-regions extends to the BOX, and does not increase in volume beyond that point once this depth is reached. Thus, even in PD-SOIs the maximum junction-capacitance of the source-to-body, C_{jSB}, and drain-to-body, C_{jDB}, is limited to a smaller value than that existing a comparably-sized MOSFET in bulk-silicon. The wiring-capacitance (i.e., associated with interconnect lines running over the field regions of an IC) is also slightly reduced by the presence of the BOX in SOI wafers.

4. Latchup Elimination. As noted in Sect. 11.1.2, the phenomenon of latchup is eliminated in CMOS-on-SOI ICs and the techniques needed to suppress latchup in bulk CMOS are unnecessary. This also means that the use of epitaxial-wafers with a heavily-doped-substrate to control latchup in some applications is no longer needed.

5. Reduction of Interwell-Isolation-Spacing. Due to the elimination of the possibility of latchup, and the ability to retain adequately low intra-well and inter-well leakage-currents due to the dielectric isolation, it is possible to reduce the isolation-spacing design rules, compared to those needed in

CMOS ICs in bulk-silicon (especially the n^+-p^+ spacing). In addition, the contact-overlap-spacing can be smaller. Thus, the packing density of the CMOS SOI-IC is increased, allowing circuits with greater functional-density to be fabricated on the same area of silicon. There is also no need for high-energy implantation of well-dopants.

6. **Reduced Soft-Error-Sensitivity.** SOI reduces the substrate-collection-volume for alpha-particle- and cosmic-ray-generated electron-hole tracks, hence lowering soft-error-sensitivity. In particular, SOI- technology for ECL-bipolar/CMOS ICs is preferred because of the high sensitivity of the ECL peripheral-circuitry to radiation.[8]

7. **Improved MOSFET Turn-Off Characteristic (Steeper Subthreshold-Slope).** The subthreshold slope, S_t, of MOSFETs built on SOI-wafers is much steeper than that of MOSFETs in bulk-Si-wafers. Subthreshold-slopes as small as 62-mV/decade have been observed in SOI MOSFETs.[9] The small value of S_t implies that an SOI-MOSFET can be turned-off with a smaller gate-voltage swing. This allows faster switching in MOSFETs, along with the use of lower threshold-voltages. A lower threshold-voltage allows lower power-supply-voltages to be used, and an increase in the drive-currents of SOI-MOSFETs compared to MOSFETs built in bulk-silicon.

8. **Reduced Leakage-Current Arising from Carrier-Generation in Junction Depletion Regions.** The ability to make the source/drain junction-areas of SOI-MOSFETs smaller by approximately two decades, also decreases the volume of the depletion-regions of these junctions. Thus, less leakage - current arising from the generation of carriers in these regions is produced, which reduces the standby leakage-current of SOI-ICs to a comparable degree. The negligible leakage current makes SOI ideal for low-power ICs.

9. **Compatibility with Bulk-Si CMOS Process Tools & Processes.** The SOI process-technology is compatible with the conventional process-technology used to fabricate CMOS ICs in bulk-silicon. Thus, the tools and facilities used to fabricate CMOS-on-bulk-silicon can also be used to fabricate CMOS-on-SOI. The SOI process-sequence is even somewhat simpler than that used to fabricate CMOS on bulk-silicon. That is, there is no need for the well-formation process in SOI-ICs, nor any need for channel-stop-implants to suppress lateral-surface-leakage between devices due to field-inversion.

10. **Advantages for System-On-Chip (SOC).** A high-resistivity substrate under the BOX can be used, which can significantly reduce the device

capacitance and the loss of passive components, such as inductors and transmission-lines, commonly used in analog- and rf-devices.[10] Therefore, SOI offers advantages over bulk-Si to realize a system-on-a-chip by integrating high-performance digital and analog/rf circuits on the same chip.

11. Increased Drain-Current. The drain current of SOI-MOSFETs, I_D, is also increased due to the significantly lower doping that can be used in the channel. The ability to tolerate lower-channel-doping occurs because of the natural punchthrough suppression arising from the backgating effects of the insulator interface.

12. Improved Circuit Reliability Due to Elimination of Junction-Spiking. When contact is made to a MOSFET in bulk-CMOS, unwanted reactions between the metal or the silicide can take place, leading to the problem of "junction-spiking" (and the attendant rise in uncontrolled leakage currents). While barrier materials normally control this problem, the potential for its occurrence is still there in bulk-CMOS. If MOSFETs are fabricated in thin SOI material, the source and drain junctions extend to the buried insulator. In that case, no leakage will be produced if some uncontrolled metal-silicon reaction occurs.

11.3.1 The Path of Using SOI May Allow Extension of the ITRS Roadmap

Many of the challenges for CMOS transistors in bulk-silicon discussed in Sect. 11.2 are relaxed or removed when SOI-wafers are used instead of bulk-silicon wafers, and SOI technologies offer new options for transistor design. With the 20-50% increase in transistor speed that can be achieved by switching to SOI, the performance of ICs can be increased without the cost of a shift to a higher lithography resolution. SOI "thin-body" CMOS transistors provide excellent control for the off-current (leakage) characteristics that limit the endurance of battery-operated computer and communication devices. Furthermore, soft-error-immunity and radiation-hardness of SOI devices is seen as a key requirement for expansion of high-data-rate network-servers that form the core of Internet and other global data-transfer-links.

In addition, SOI-wafers are opening a path for the extension of planar CMOS transistor design by perhaps a decade beyond the "end-of-roadmap" limits foreseen in the ITRS-2001 study for 50-nm gates on bulk-silicon. Operational SOI transistors have been studied with gate lengths as small as 25-nm and Si channel thicknesses of 1-nm.

TABLE 11-2 COMPARISON OF KEY ISSUES FOR SUB-100-nm CMOS FOR BULK-SILICON AND FULLY-DEPLETED SOI MOSFETS

Critical Issue	Bulk silicon	SOI
Junction depth	RTP/(LTP)	Set by Si-SOI film
Dopant activation	Thermal budget on vertical and lateral diffusion	Lateral diffusion limit
Junction abruptness	Critical vertical and lateral abruptness limits	Carrier injection still requires abrupt lateral junction in thin-body
Contacts	Extended S/D	Extended S/D
Off-current	Requires halo	Dual gate
Threshold control	SSRD (super-steep retrograde doped channels) or SiGe gate	SiGe gate
Well doping	High-energy implant	No well

To begin, it is easier to control the junction-depths of the source/drain extensions in SOI devices. In bulk-silicon, obtaining shallow-junctions requires careful control of thermal budgets, with the use of rapid-thermal-annealing-"spikes" or laser-annealing to limit both the vertical- and lateral-dopant-diffusion, while obtaining high degrees of dopant-activation and adequate defect annealing. With fully-depleted, thin SOI (FD-SOI) MOSFETs (discussed in Sect. 11.5.2), the source/drain-extension junction-depth is determined by the Si-SOI-layer thickness. This means that the thermal-budget-constraints on vertical-junction-depth are essentially removed for fully-depleted-SOI CMOS devices. Second, the ability to activate dopants in the source/drain-regions of CMOS in bulk-silicon is limited by the thermal-budget on vertical- and lateral-diffusion effects, while in FD-SOI MOSFETs only the lateral-diffusion sets a limit on this thermal-budget (a potentially less difficult problem to solve). Third, in CMOS-on bulk-silicon both the vertical- and lateral-abruptness of the source/drain-extension doping-profile needs to be controlled, whereas in FD-SOI, only the abruptness of the lateral-doping-profile of the source/drain extension-region needs to be controlled. Fourth, the well-doping in CMOS-on-bulk-silicon needs to be carried out with high-energy ion-implanters, but in CMOS-on-SOI, no wells even need to be formed. Finally, SOI-technologies can be developed to include new device structures, such as dual-gate and high-channel-mobility (with

Si:Ge-CMOS), that provide new tools for transistor designers to increase signal speed while decreasing off-current leakage-levels. The transistor characteristics and process tradeoffs for bulk-silicon- and SOI-wafers are summarized in Table 11-2. The process challenges shift to controlling lateral-diffusion and obtaining abrupt-junction-interfaces.

One of the critical processes that will be needed to allow thin-SOI-wafers to serve as the substrates for sub-50-nm CMOS is the ability to fabricate thin-SOI wafers with high-levels of control over the SOI-Si device-layer-thickness, electrical-properties, and interface-characteristics. The Smart-Cut® and Nanocleave® processes described in Sects. 11.6.3.2 and 11.6.3.3, respectively, are two approaches that may allow these goals to be met.

The timeline for the use of the various types of "thin" SOI MOSFETs can be linked to the projections made by the ITRS-2001 (Fig. 11-7). The thickness of the Si-to-SOI surface layer (T_{Si}) will shrink according to the approximate design rules of $T_{Si} \sim 2 \times X_{j(S/D\text{contact region})}$ for partially depleted SOI MOSFETs (see Fig. 11-2a, which shows what is meant by the terms *"S/D contact-region"* and *"S/D extension region"*), and $T_{Si} \sim 1 \times X_{j(S/D\text{extension})}$. Up to the year 2000 timeframe, production SOI-ICs were being manufactured using PD-SOI.

For the time between 2001 to 2005 (i.e., for MOSFETs with gate lengths of 0.15-μm down to 0.07-μm), the Si-SOI layer thicknesses being used is expected to cause the type of SOI FETs to be a mix of partially-depleted (PD) CMOS for mainstream logic and ASIC devices, and fully-depleted (FD) SOI for high-performance processor and server/router devices.

For CMOS devices with gate-lengths in the range of 50-nm (beginning in about 2006), the multitude of process and performance issues described in the ITRS study will preclude the use of planar-CMOS on bulk-Si as well as in partially-depleted SOI. By migrating mainstream transistor structures to fully-depleted SOI, it may also be possible to extend CMOS circuit technologies a full decade beyond the "end-of-roadmap" for planar-CMOS on bulk-silicon. "Thin-body" SOI, with gates sizes at 20-nm on Si-SOI layers of ~10-nm are now at the threshold of research devices. If scaling continues at the historical rate of a 0.7 shrink every three years, these devices would be in production by about 2015.

11.4 HISTORY OF SILICON-ON-INSULATOR (SOI) TECHNOLOGY

Silicon-on-insulator technology has been available since the late 1960's, and has been used since that time for military integrated-circuit applications. However, it has taken more than 30 years for the first mainstream commercial SOI devices

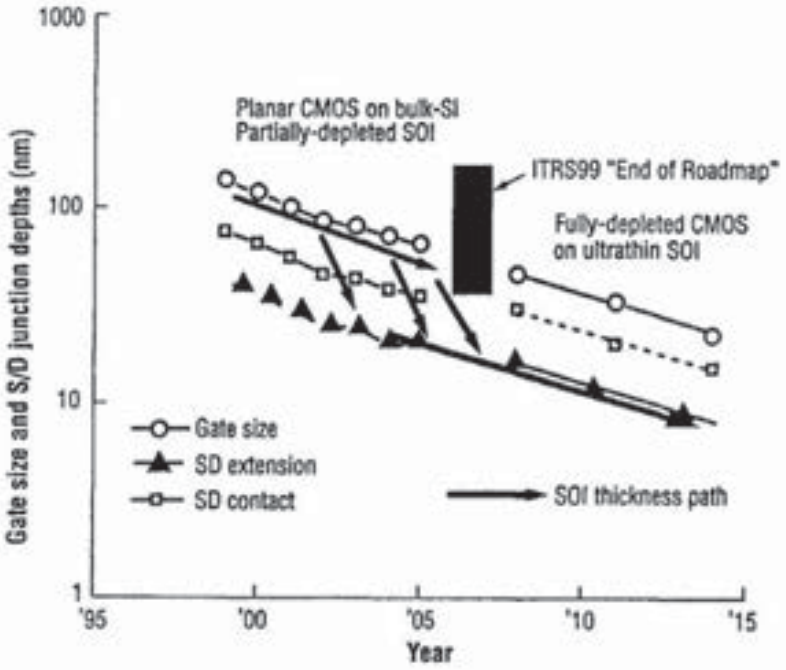


Fig. 11-7 Timeline of Si-SOI layer thickness linked to junction depth and gate size predictions from ITRS-2001.

to be offered. As described in Sect. 11.3, the advantages of SOI technology are well known. First, circuits fabricated in SOI wafers have reduced parasitic-capacitance when compared to bulk or epi wafers. Less capacitance translates to lower power or higher speed. Second, SOI devices have improved radiation-induced single-event-upset (SEU) immunity, and thus they are useful for space applications. Even large memory arrays that are used in terrestrial applications are subject to SEU and may require SOI circuits. Third, SOI devices are completely free of latchup. Finally, the fabrication process in SOI wafers can be simplified by reducing the number of masks by as much as 30%. However, SOI wafers are more expensive to fabricate and may contain more defects than bulk or epi-silicon. As the requirements for improved speed, reduced power, and less expensive processing have become more important, a keen interest in SOI technology has been re-kindled among mainstream device suppliers.[41]

Currently there are a number of useful approaches for making SOI wafers. The first group includes the "legacy" SOI technologies of *dielectric isolation* (see Vol. 2, Chap. 2, for more information on the dielectric-isolation-technology) and *silicon-on-sapphire* (*SOS*) that were developed in the 1960's, and that even today, still find limited application. (As we will note, an improved SOS technology is currently being used to fabricate high-performance rf-devices.) The more modern SOI-technologies are based on either a buried-insulator-layer

created by ion-implantation technology (e.g., *Separation by IMplanted OXygen* [*SIMOX*]), or by a *wafer bonding* (*WB*) technique. (The insulating substrate for SOS wafers is sapphire [i.e., very pure single-crystal, aluminum oxide, Al_2O_3]. On the other hand, the insulator for the SIMOX and wafer-bonding technologies is silicon dioxide.) Various (and to-date, unsuccessful) attempts have been made to grow single-crystal silicon (or large-grain polycrystalline-silicon) directly on amorphous insulators.[42] When devices were first built on SOI-wafers, small islands of silicon were formed (typically by etch techniques) on top of the insulator. In modern SOI ICs, either LOCOS or STI is used to complete the dielectric isolation. Individual devices are then fashioned in the islands and these devices are then interconnected in a conventional manner. SOI-technology has been the subject of intense research and development in the late 1990s. A number of symposia that address SOI technology are held on a regular basis.[44,45] For more information on the evolution of SOI circuits, see Sect. 11.8.

11.5 SILICON-ON-INSULATOR DEVICES

The MOSFETs built using SOI-substrates are categorized in a number of ways. The first such grouping differentiates between so-called *thick-SOI-MOSFETs* (built on SOI-substrates whose Si layer is thicker than 1 micron - i.e., > 1.0-μm), and *thin-SOI-MOSFETs* (in which this layer is less than 1 micron thick – i.e., < 1.0-μm). The applications for "thick" SOI include power-IC switches, rf bipolar-signal-processors, and a wide variety of MEMS and optoelectronic devices. Thick-SOI devices were also historically used for a variety of military applications (e.g., radiation-hardened ICs) and for such commercial-applications as high-voltage and high-power devices. The principal SOI-material being presently used for VLSI and ULSI CMOS signal, logic, and memory applications is "thin" SOI (50-1000-nm thick [0.005-1-μm]. The "thick" SOI market segment, at only $50 million/year is a very small piece of the $110 billion/year commercial, mainstream CMOS IC market (and is not likely to grow significantly in the future). Hence, we will not further deal with such devices, but instead will focus on "thin" SOI devices in the rest of this section.

Within the category of "thin" SOI devices, there are a number of additional subcategories, namely *partially-depleted SOI-MOSFETs* and *fully-depleted SOI-MOSFETs*. The thickness of the silicon-device-layer of the SOI-structure can be chosen so that the SOI-MOSFETs will operate in one or the other of these modes. Each of these SOI device types will be discussed separately in more detail. Note once again that one of the ways that MOSFETs fabricated in SOI differ from the MOSFETs fabricated in a bulk (i.e., non-SOI) substrate is that the

channel region (or body) of the SOI-MOSFET is not automatically connected to any external terminal (i.e., intrinsically the body of the SOI MOSFET is electrically floating). This aspect of SOI-MOSFETs (a floating body), leads to so-called *floating-body-effects,* which can be detrimental to device performance. Floating-body effects in SOI-MOSFETs will be described in more detail later.

11.5.1 Partially-Depleted Thin-SOI-MOSFETs (PD-SOI)

Partially-depleted, thin-SOI-MOSFETs (PD-SOIs) are given this name because when the maximum gate-bias is applied to the MOSFET, the depth of the resultant depletion-region in the channel of such devices is shallower than the thickness of the silicon layer (Fig. 11-8). Note that "thick" SOI MOSFETs are also operated as partially-depleted MOSFETs.

As mentioned earlier, the body of SOI-MOSFETs is intrinsically floating. That is, unless a contact to some circuit potential is explicitly provided, the potential of the SOI-body is floating (Fig. 11-9a). However, if the SOI-MOSFET body is grounded (i.e., by independent body contacts or by body-source ties), partially depleted SOI MOSFETs behave very much like bulk-Si MOSFETs, and most of the standard $I_D(V_{GS},V_{DS})$ equations and design concepts apply. If body-contacts are not supplied, so-called *floating-body effects* arise, leading to potentially detrimental consequences (including the introduction of a *kink* in the dc I-V characteristics, a reduction in breakdown-voltage, and the causing of hysteresis and instability during dynamic operations). The issue of floating-body

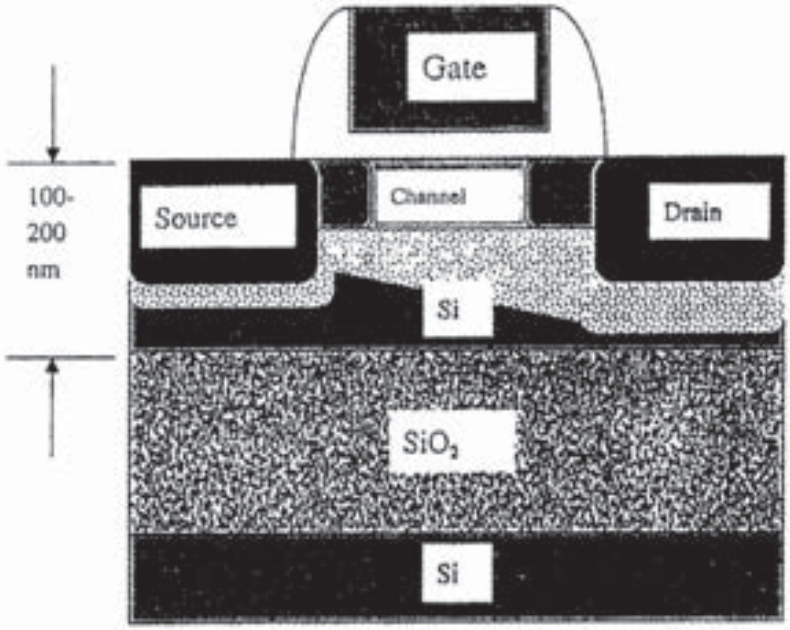


Fig. 11-8 Partially-depleted transistor in a Si layer over an isolation layer.

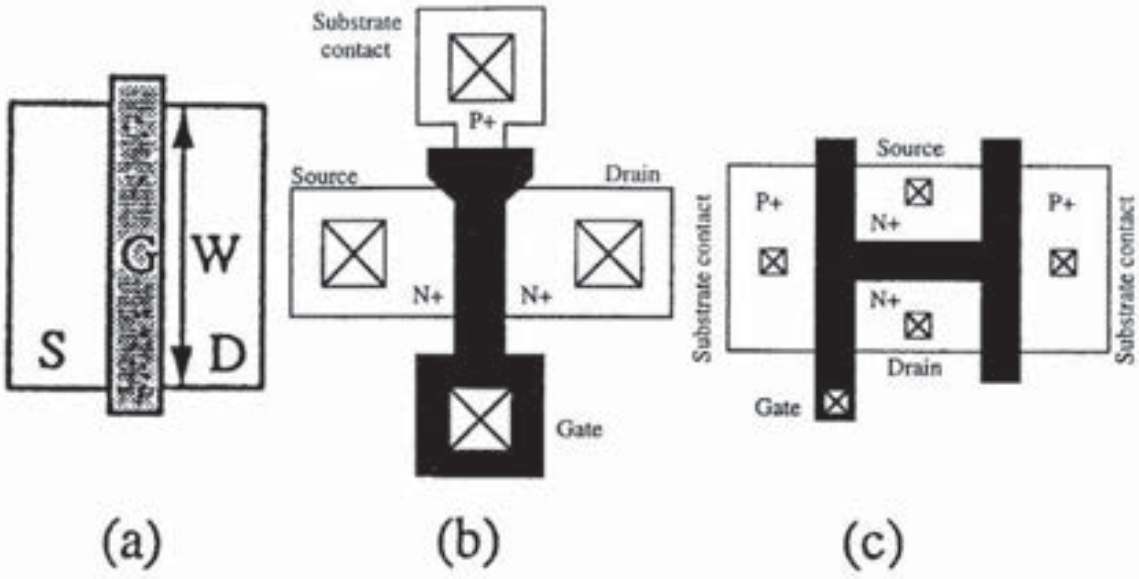


Fig. 11-9 a) Top-down view of an SOI-MOSFET in which no provision is made for contacting the body. Unless a contact to some circuit potential is explicitly provided, the potential of the SOI-body is floating; b) MOSFET with body-contact made on one side of the MOSFET channel; c) An "H-gate" SOI-MOSFET (i.e., a body-contact is made on each side of the MOSFET channel).

effects in partially depleted SOI MOSFETs is described in further in Sect. 11.9.

Note that body- and substrate-contacts were routinely used when "thick" SOI devices were fabricated, as shown in Fig. 11-9b. To avoid the possible device-performance-degradation that floating-body-effects could cause, early partially-depleted, thin-SOI-MOSFETs also used either separate body-contacts or source-body ties, again as shown in Fig. 11-9b. However, since the body-region of the MOSFET beneath the channel can exhibit a high resistance, a single body-contact tied to ground may not be adequate. In some applications a body contact is made on both sides of the MOSFET channel, forming a so-called "H-gate" (Fig. 11-9c).[11] Figure 11-10 shows how a body-contact is made to an SOI-NMOSFET both from a top view, and from a cross-sectional view. Figure 11-11 shows the layout of a DRAM sense-amplifier-circuit fabricated in SOI, in which a body-to-source-tie is employed.

While body-contacts (and body-source-ties) can be used to avoid the problems associated with the floating-body-effect, these structures consume valuable silicon-circuit-area. That is, the global use of body-contacts on every SOI MOSFET in a circuit may severely degrade the density (and hence also the

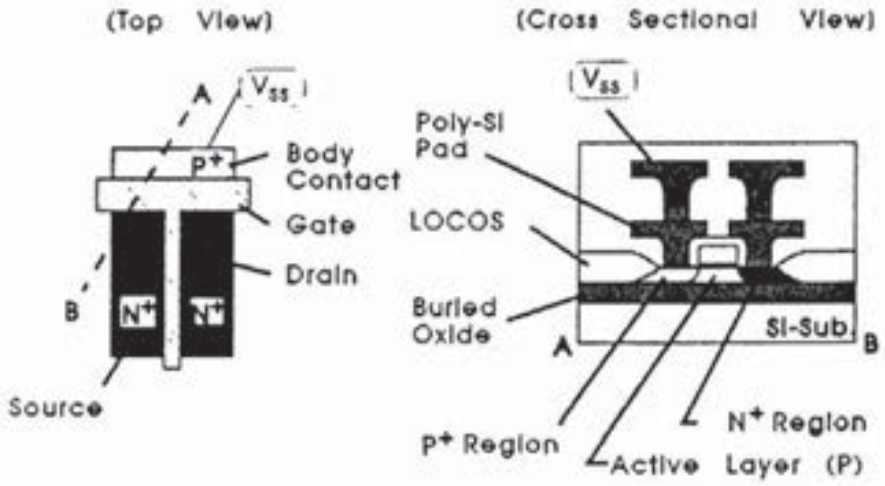


Fig. 11-10 A top-down and diagonal cross-sectional view of an SOI-NMOSFET with a body-contact made on one side of the MOSFET channel.

performance), of a large-scale IC design. As a result, in more recent SOI ICs built using partially-depleted thin SOI MOSFETs, attempts were made to find ways to eliminate the need to use such body-contacts. As shall be seen in Sect. 11.9, it has been demonstrated that in some circuit applications, such body-contacts may not need to used. That is, the device-performance-degradation caused by floating-body-effects may be small enough to allow the elimination of body-contacts from such SOI circuits.

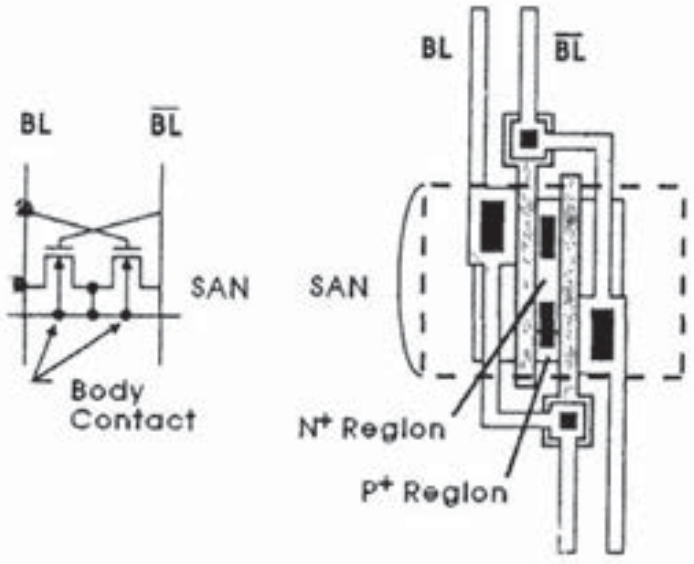


Fig. 11-11 Layout of a SOI-NMOSFET DRAM sense amplifier incorporating a body contact.

When CMOS circuits are built with partially-depleted thin-SOI-MOSFETs (and are powered with the same voltages), switching-speeds are observed to increase by 20-30% compared to equivalent circuits in bulk-silicon devices. Alternatively, partially-depleted SOI devices can be operated at lower voltages and lower battery-power-drains with a significantly-smaller speed-penalty than bulk-silicon devices. As noted earlier, since partially-depleted SOI-MOSFETs operate electrically in essentially the same manner as bulk MOSFETs, an immediately useful advantage is that circuit designs from bulk-silicon devices can be imported into the SOI environment with only a modest investment in circuit redesign. As of 2000, the SOI circuits being offered commercially were largely based on partially-depleted SOI-MOSFETs.

11.5.2 Fully-Depleted Thin SOI MOSFETs (FD-SOIs)

In *fully-depleted, thin-SOI-MOSFETs* (FD-SOIs) the depth of the depletion-region in the channel of the transistor extends entirely through the thickness of the surface-silicon-layer of the SOI-structure (Fig. 11-13b), even when there is no-bias applied to the gate. (In addition, the junction-depth formed in the source/drain-regions in fully-depleted SOI-structures is also as deep as the Si-surface-layer-thickness. Hence, there is no *pn*-junction at the [source/drain]/BOX interface, and thus there is no depletion-region at that interface to contribute to the source/body- and drain/body-capacitance.) This also implies that there is less overall junction-surface area than in a CMOS-structure formed in bulk-silicon (which results in reduced leakage-currents and junction-capacitances). In the FD-SOI case, the depletion-charge is thus constant and cannot extend further into the body as the gate-bias is increased. As a result, the application of a gate-bias results in more inversion-layer-charge for the same gate-bias as in a bulk-Si MOSFET. This has several beneficial effects, including: **1**) drain current is increased; and **2**) subthreshold-slope becomes steeper (Fig. 11-12). However, the $I_D(V_{GS},V_{DS})$ equations used to model bulk MOSFET and PD-SOI MOSFETS are no longer valid, and new I-V relationships need to be used for FD-SOIs. Although circuit-designs need to take full account of the special conditions of operation in an SOI layer, the highest gains in circuit-speed, reduced-power requirements, and the highest level of soft-error-immunity are realized with fully-depleted SOI-devices. That is, FD-SOI devices exhibit the following advantages over bulk and partially-depleted SOI-MOSFETs: **1**) sharper subthreshold slope;[12] **2**) higher drive-current; **3**) higher-transconductance; **4**) reduced parasitic-capacitance; and **5**) reduced short-channel-effects.[13] The latter benefit is gained because some of the charge that would otherwise be needed to support the

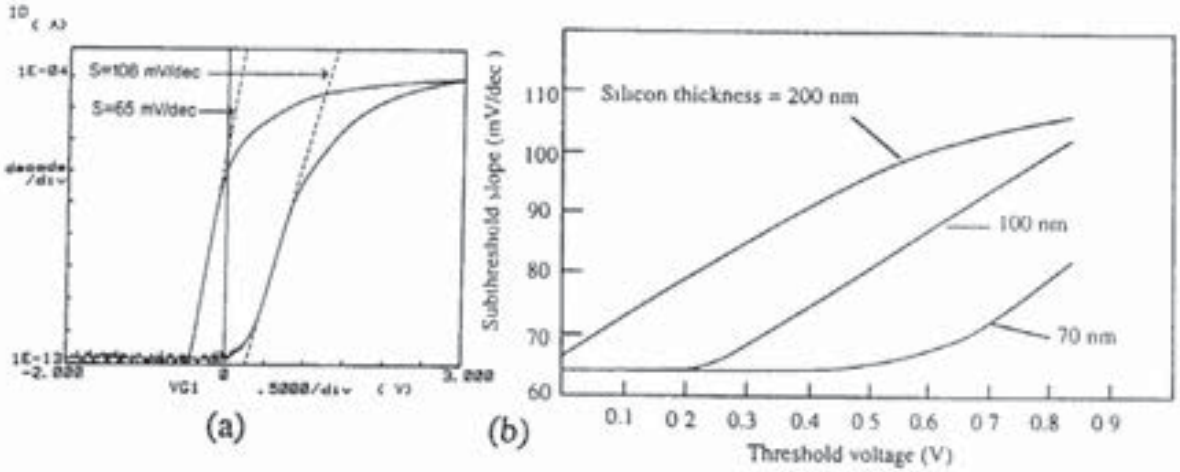


Fig. 11-12 a) Log-plot of drain-current I_D versus gate-voltage in SOI-MOSFETs made with two different silicon film thicknesses: 90-nm (left curve) and 450-nm (right curve). The inverse-subthreshold-slope, S_t (in millivolts per decade) is indicated. V_{DS} = 1 V; b) S_t as a function of threshold voltage for various SOI film thicknesses (*n*-channel device).

channel depletion-region-charge is instead produced by charge in the substrate under the BOX. However, as of 2001, fully-depleted SOI MOSFETs were not yet being widely implemented in commercially available ICs.

The reasons fully-depleted SOI MOSFETs were not being used in commercial production by 2001 is that these devices must meet the requirement that the silicon-film-thickness must always remain well below the depletion-region-width in the channel, during all modes of the MOSFET operation (i.e., from when the device is OFF through inversion). This dictates the use of a low threshold-voltage-value, and the ability to keep the Si-SOI surface-layer-thickness well below the depletion-region-width under conditions of thickness variation. That is, it must be possible to control the Si-surface-layer-thickness to within 5-10-nm.[14] This imposes severe limitations on the manufacturability of the device. (Use of partially-depleted SOI MOSFETs in an SOI-circuit, on the other hand, alleviates the constraint on V_T and its sensitivity, allowing the channel-doping-profile to be tailored to meet any desired V_T value, thus easing the manufacturing problem.) Nevertheless, by the end of 2001 this situation was starting to change. Intel announced a 50-nm FD-SOI structure, and said that they would use FD-SOI MOSFETs to build ICs from 2005 and beyond.[66] In addition to FD-SOI, the Intel MOSFETs incorporate elevated source/drains (see Sect. 5.3.2).

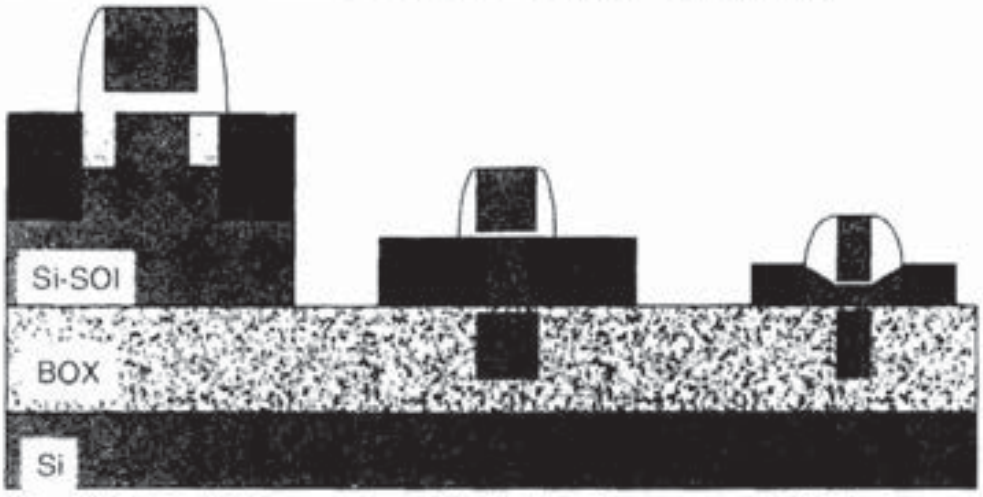


Fig. 11-13 Schematics of partially-, fully-depleted, and "thin-body" (double gate) SOI MOSFETs.

In FD-SOIs some of the gate-charge is balanced by charge in the depletion-region of the channel, but additional charge will arise in the substrate beneath the BOX layer. This induced charge in the substrate will have the effect of trying to establish an inversion layer *along the interface between the Si surface layer and the BOX* (i.e., the substrate beneath the BOX acts as a *back-gate*). As a result, a coupling exists between the potentials on the top-gate and the back-gate, and the electrical characteristics of the channel along the back-gate vary with bias applied to the top-gate. For example, leakage along the back-gate-channel will be caused by the application of a bias on the top-gate. In fact, when the Si-surface-layer of an FD-SOI becomes too thin, the top-gate eventually can no adequately longer turn-off the FD-SOI MOSFET. In such cases it may be necessary to fabricate these FD-SOI MOSFETs with a dual-gate structure and a thinned body-region (as discussed in Sect. 11.5.3.2 and as shown schematically in Fig. 11-13c). Such dual-gate FD-SOI-devices offer the prospect of excellent control of the off-current of small channel-length SOI-MOSFETs.

A dual-gate FD-SOI has both a top-gate and bottom-gate (Fig. 11-13c). The channel is sufficiently thin so as to be fully-depleted and an inversion-layer is established at both the top and bottom gates. To minimize short-channel-effects the channel-thickness must be considerably smaller than the gate-length.[48] For gate lengths smaller than 50-nm, this will require thin silicon layers of 20-nm or

less. The potential transconductance of a dual-gate device is approximately twice that of a bulk-device because of the presence of two gates. Also, the subthreshold slope, S_t, is potentially much improved over a conventional MOS device. Several theoretical studies have shown that the silicon thickness for a single-gate FD-SOI device must be about 1/2 that of a dual-gate device for equivalent short-channel effects.[48]

However, by 2000, dual-gate FD-SOIs were still in their infancy, and were far from being ready for consideration as devices in commercial ICs. That is, there are major problems in the practical manufacturing of dual-gate FD-SOIs. In a horizontal configuration, a gate must be fabricated below a thin silicon layer and accurate alignment of the top-gate over the bottom-gate must be achieved. All of these steps are very difficult.

11.5.3 Process Technologies Used to Fabricate Fully-Depleted SOI Devices

As noted earlier, one reason it is difficult to manufacture FD-SOI is that it must be possible to fabricate very thin Si-SOI surface-layers and to control their thickness to within 5-10-nm. Here some of the process sequences that have been developed to try to meet these requirements are listed.

11.5.3.1 FD-SOI Fabrication by Thinning the Si-SOI Surface Layer by Thermal Oxidation: An early process for fabricating FD-SOI devices was reported in 1992.[15] In this work a 220-nm-thick Si-SOI layer formed with a SIMOX process was the starting substrate. This layer was thinned to a 40-nm-thickness by carrying out a blanket thermal-oxidation of the Si-SOI layer and then etching off the oxide in an HF etch step. A LOCOS isolation-structure was used to provide lateral isolation. After forming the gate, a titanium salicide process was used to form $TiSi_2$ on the source-drain regions and on the polysilicon gate. The silicon in the source-drain regions was completely converted to $TiSi_2$ so that the silicide layer lies on top of the buried-oxide. A silicidation that fully consumes the SOI layer in the source/drain regions was chosen for ease of control. Figure 11-14 shows a cross-section of the finished device.

11.5.3.2 FD-SOI Fabrication by a LOCOS-Recess of the MOSFET Channel-Region: As the thickness of the Si-SOI layer (T_{Si}) has to become even thinner than the 40-nm-thick Si layer used in the FD-SOI device of the previous section, the parasitic resistance of the source/drain regions of such devices becomes too large to obtain the required device performance characteristics (even if these regions

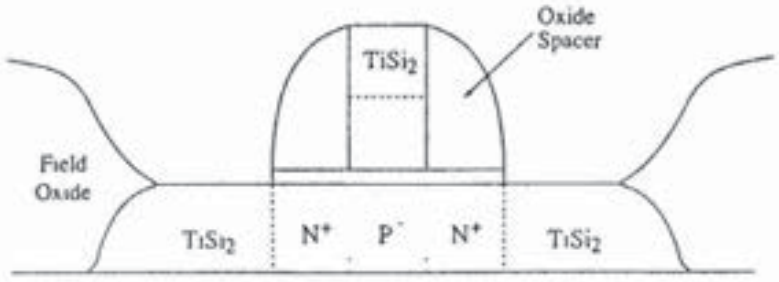


Fig. 11-14 Cross-section of the ultrathin SOI transistor with silicided source/drain and gate.

are silicided). An approach which allows the S/D regions to remain thicker while reducing the thickness of the body has thus been developed. In this method, the body region of the FD-SOI device is thinned by using LOCOS process, as shown in Fig. 11-15.[16] In one example, a SOI wafer with a T_{Si} of 195-nm was the starting material. A LOCOS process was used to thin the body region of the FD-SOI device to a 110-nm thickness. MESA active islands were formed by selectively etching back the Si-SOI to the BOX layer. These MESA islands were then oxidized to provide the lateral isolation.[17] Figure 11-16 shows an SEM of a completely fabricated FD-SOI device.

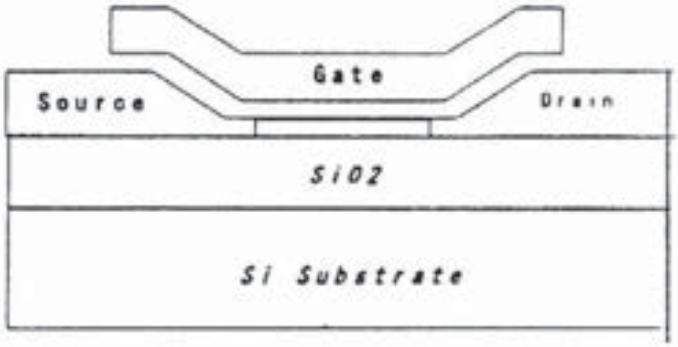


Fig. 11-15 Cross-section of a LOCOS-thinned "thin-body" SOI MOSFET.

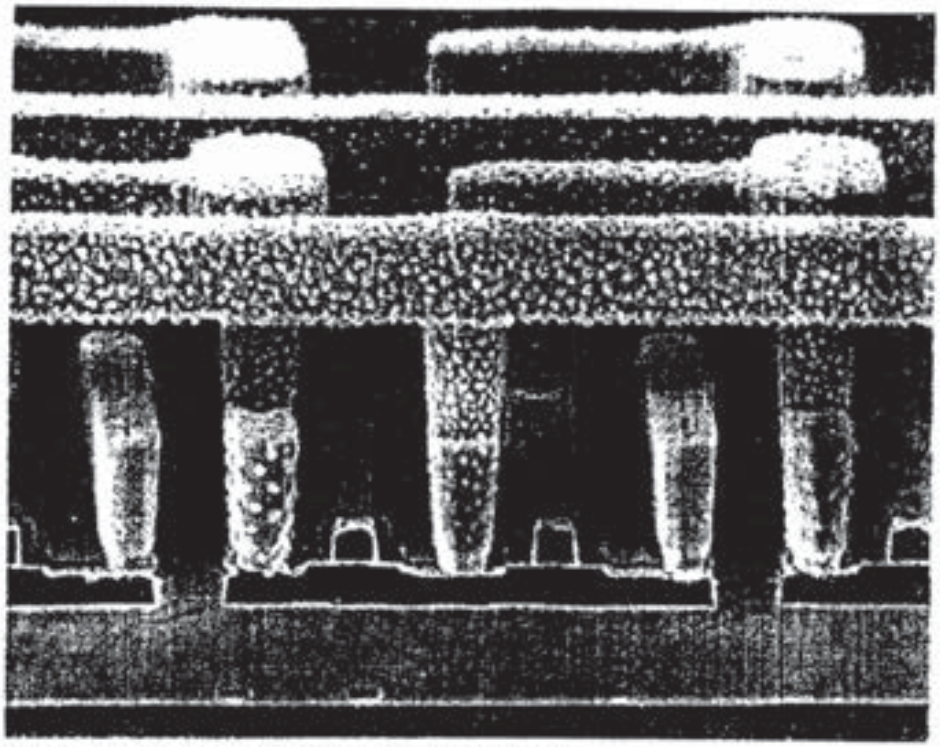


Fig. 11-16 SEM of completely fabricated SOI-MOSFET.

11.6 FABRICATION TECHNOLOGIES USED TO PRODUCE STARTING SOI-WAFERS

There a number of methods for fabricating starting SOI-wafers. Here those being currently used to make SOI wafers will be discussed. The annual production of SOI wafers in 1998 was only 10-million square-inches of silicon wafer. However, it is predicted that the total SOI-wafer-demand will grow by a compound annual growth rate (CAGR) of 55% between 1999 and 2005. In 2005, this should make the production of SOI wafers reach 187-million square-inches. The SOI merchant-wafer market registered $100 in revenue in 1999 (this did not include the captive SOI wafer production by such companies as IBM).

11.6.1 Silicon-on-Sapphire (SOS)

The work on silicon-on-sapphire (SOS) dates back to the 1960s when there was interest in this technology for its well-known immunity to radiation damage. Sapphire is insulating, and so circuits built in SOS substrates have high speed and excellent device-isolation. That is, unlike in other SOI approaches (where a buried oxide-layer is used to provide vertical isolation), the entire sapphire substrate is an insulating material. This reduces microwave losses and makes SOS an ideal candidate for the integration of rf-circuits for wireless

communications. Furthermore, the high thermal-conductivity of sapphire facilitates thermal dissipation through the substrate, which alleviates the self-heating effects of the other SOI technologies as described in Sect. 11.9. Device isolation can be easily achieved through simply removing the conducting silicon between the transistors by etching down to the insulating sapphire-substrate. SOS circuits were historically used for radiation-hardened applications. There are several excellent review articles on this subject and the reader is referred to them for additional details.[18,19]

Since SOS technology is limited to fabrication of MOS devices, a (100) orientation of the crystalline-silicon film is desirable. This orientation has been hetero-epitaxially grown on several different sapphire (i.e., crystalline alumina) orientations, specifically (0112), (1012), and (1102). The lattice-parameter mismatch between the Si-epi layer and the sapphire-substrate is sufficiently large to cause the generation of misfit dislocations, stacking faults, and edge dislocations in the Si-epi layer. The defect-concentration is heaviest near the substrate/epi interface and decreases as the film grows.[20] Figure 11-17 shows the stacking fault density as a function of epitaxial-film-thickness in a thick SOS film. However, current requirements for fabrication employ so-called *fully-depleted MOSFETs* (which are defined in Sect. 11.5) and these must be built in a Si-epi films only 0.1–0.3-μm thick. These films can be highly faulted, and require additional processing to improve their crystalline quality.

The accepted way to improve the film quality is the technique of amorphization (by implanting a high dose of Si into the deposited silicon

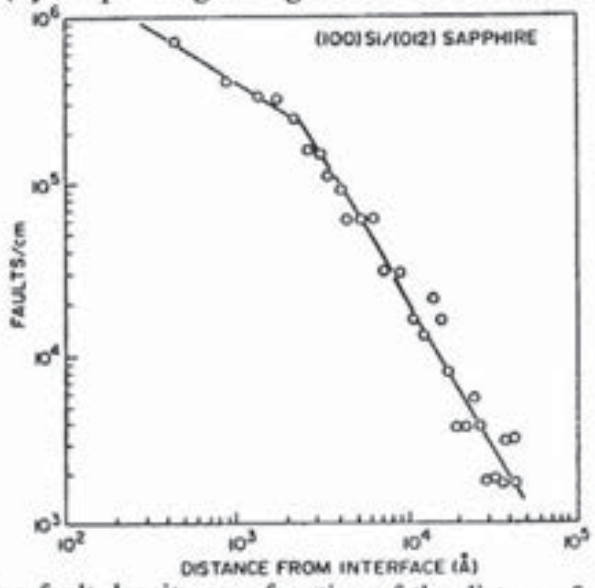


Fig. 11-17 Stacking-fault-density as a function of the distance from the silicon/sapphire interface in an MOS structure.[21] Reprinted by permission of the American Physical Soc.

epitaxial layer) followed by *solid-phase-epitaxial regrowth (SPE)*.[22] In *thick SOS films* only a single amorphization is required. That is, an ion implantation step causes the region near the interface to become amorphized. Then the undamaged single-crystal-Si near the surface is used as seed material for the epitaxial regrowth of the amorphous layer. Since the regrowth is primarily along the <100> direction, stacking faults and twins do not propagate, leaving a controllable number of dislocations in the film.

For *thinner SOS films* a *double SPE (DSPE)* process may be required. In this approach, the Si-region near the sapphire-interface is first amorphized by a high-energy ion-implantation step, and the surface-regions are used as a seed for regrowth.[23] This is followed by a lower-energy ion-implantation step that then amorphizes the surface-region of the Si. The previously re-crystallized bottom layer is now utilized as the seed to re-grow the surface-amorphous-layer. Devices fabricated in such material show higher surface mobilities and lower junction leakage than either untreated SOS, or material obtained from a single SPE-process.[24] Peregrine Semiconductor, which markets high-performance SOS-rf devices for the commercial market, builds these devices using standard ULSI processing techniques on 200-mm SOS wafers. SOS films may require additional characterization over that used for conventional epi-films. These methods are comprehensively reviewed by Cullen *et al*,[25] and Munteanu *et al*.[26]

11.6.2 SOI by Separation by Implanted Oxygen (SIMOX)

A widely used approach to manufacture SOI wafers is that of *Separation by IMplanted Oxygen*, or SIMOX. This process was first proposed by Izumi of NTT in 1978 (and who also coined the acronym SIMOX).[27] The SIMOX technology

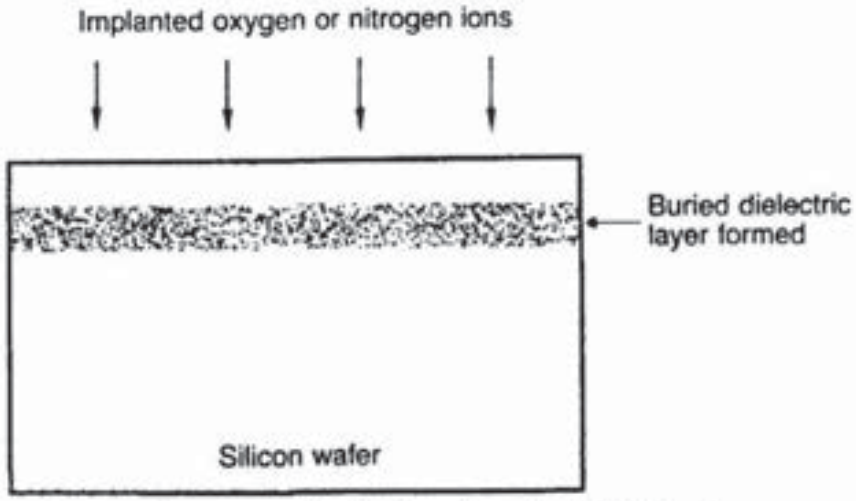


Fig. 11-18 Implanted oxygen ion process for the formation of a SIMOX wafer.

is based on the ion-implanting a high-dose of oxygen ions (O^+) into a silicon wafer (~$1.4x10^{18}$ O^+ ions/cm^2), such that the peak concentration lies about a half micron beneath the silicon surface. After implantation these wafers are subjected to a high-temperature-anneal which forms a continuous stoichiometric subsurface-layer of silicon-dioxide (Fig. 11-18). This *buried-oxide* (or BOX) layer electrically separates the top and bottom portions of the wafer, and is typically 350-nm thick. The single-crystal-region of silicon above the BOX is typically 250-nm thick. The result is a wafer with a thin surface-silicon-film of high-crystalline-quality, which also has an abrupt interface to a buried layer of stoichiometric-SiO_2.

The SIMOX process has undergone significant evolution since it was first introduced. In early-SIMOX the material was of very-poor quality. That is, the defect-density (dislocations, oxide-precipitates, and polycrystalline-silicon-inclusions) was so high that epitaxial-silicon had to be grown on top of the silicon overlayer in order to obtain a silicon film good enough for device fabrication. In addition, at that time, the implantation of sufficient oxygen to form a buried-oxide-layer in a 2x2 cm^2 area on a silicon-wafer required 24 hours (using a conventional implanter). Thus, most SOI researchers viewed SIMOX as just a laboratory curiosity. Nevertheless, when NTT produced a 1-k SRAM using SIMOX in 1982, the SOI community began to consider SIMOX from a less skeptical perspective.

In 1985, Eaton (now Axcelis) produced the first high-current oxygen-implanter, the NV-200, according to specifications provided by NTT. This machine could implant O^+ ions at a beam-current of 100-mA and at energies up to 200-keV. The availability of such an implanter was a key step toward allowing SIMOX development to go forward.

Before 1985, SIMOX wafers were annealed at 1150°C (which is the highest temperature available in conventional furnaces equipped with quartz tubes). In 1985, it was demonstrated that annealing at 1300°C in a furnace equipped with either a polysilicon or a silicon-carbide tube (and even at higher temperatures - up to 1405°C - in a lamp-annealing system) could dramatically improve the quality of the material, and produce atomically sharp silicon/silicon-dioxide interfaces. Other improvements (to be discussed in more detail later), included the use of multiple-implants, and a lower oxygen-dose. Eventually, by the late 1980s, SIMOX-wafers began to be produced commercially by publicly-traded companies (e.g., Ibis Technology).

As a result of all of this progress, three forms of SIMOX have been developed. The first two forms (*high-dose SIMOX* and *low-dose SIMOX*) use

conventional ion-implantation technology to implant a high-concentration of oxygen. A third version of the SIMOX process, which uses a plasma immersion implanter for the oxygen implant, is called *Separation by Plasma IMplantation of OXygen* (SPIMOX). The latter technique is a potentially low-cost process for fabricating SOI substrates, because the oxygen atoms are implanted by *plasma-immersion ion-implantation* (PIII, see Fig. 11-26). The whole wafer is implanted at once and implantation is thus independent of wafer size. This results in high wafer throughput. A possible drawback of this technique is the lack of ion-beam-selection and the fact that this process may cause more contamination in the SOI wafer than would a conventional ion-implanter.[28] High-dose SIMOX was the first technique to be developed, and it was followed more recently by the low-dose SIMOX-process. The high-dose SIMOX and low-dose SIMOX forms are described in more detail here.

Although SIMOX wafers can be obtained from several commercial suppliers, the high-dose SIMOX wafers are still significantly more costly than bulk-Si or even epi-Si wafers.[29] However, the development of the low-dose SIMOX process appears to significantly cut the premium that must be paid for SIMOX wafers. In fact, it is estimated that SIMOX wafers in the future may not cost any more than epitaxial wafers (whereas in 2000 they cost 2-3x the cost of epi wafers). Future cost-reductions in SIMOX wafers will come from such advances as increased beam-current (e.g., 2x that of present implanter-technology). Note that the SIMOX process has only two major steps: **1**) oxygen implantation; and **2**) annealing. Since SIMOX wafers are commercially available from various suppliers, introducing SIMOX wafers into production does not require any major infrastructure changes in a fab. Furthermore, after years of development work, SIMOX-SOI-wafer and device-fabrication processes have achieved sufficiently high-quality and low-cost to be applied commercially. In August 1998, IBM announced it would apply its leading-edge CMOS technology to commercialize SOI-CMOS products based on low-dose-SIMOX SOI.

11.6.2.1 High-Dose SIMOX

Figure 11-18 shows a schematic of the high-dose SIMOX process. The sequence of steps involved in the high-dose SIMOX process is as follows:

1. The starting material is typically a (100) device-quality Si-wafer. The wafer is first subjected to a high dose (1-2x10^{18}/cm^2) oxygen (O^+) ion-implantation step at high enough energy (150–300 keV) so that the peak (projected range) of the implant is deep within the silicon (0.3–0.5-μm).

This process is carried out using a dedicated high-current oxygen-ion-implanter, typically with a beam current of 100-mA. For reasons described in the next paragraph, the wafer is maintained at a temperature of ~600°C during the oxygen-implant step.

Note that stoichiometric-silicon-dioxide contains $4.4x10^{22}$ oxygen atoms/cm^3. Therefore, from such considerations alone, an implantation dose of $4.4x10^{17}/cm^2$ should be sufficient to produce a 100-nm-thick buried-oxide-layer. Unfortunately, due to the statistical nature of ion implantation, the oxygen-profile in silicon does not have a box shape, but rather a Gaussian profile. Thus, these implanted-oxygen atoms will spread out over a wider range than 100-nm, and silicon-dioxide-stoichiometry is not reached. If the wafer was annealed after such a $1\text{-}2x10^{18}/cm^2$ oxygen-dose, oxygen precipitates would be formed at a depth equal to the depth of the maximum oxygen-concentration, but no continuous layer of SiO_2 would be formed. Experiments have shown that a dose of $1.4x10^{18}$ oxygen atoms/cm^2 must be implanted in order to create a continuous-buried-oxide-layer. The standard dose (which is the most common) is $1.8x10^{18}/cm^2$, and produces a 400-nm-thick buried-oxide-layer upon annealing. As the dose is increased above $1.8x10^{18}/cm^2$, the overall oxygen-implant-profile no longer shows a higher peak value, but instead becomes broadened. This is due to the fact that at the elevated temperature of the implant, oxygen can readily diffuse to the Si/SiO_2 interface, where oxidation occurs. The dose of oxygen at which the continuous-buried-oxide-layer begins to form ($1.4x10^{18}/cm^2$) is called the "*critical dose*."

2. The wafers are then given a post-implant anneal for sufficient time (3–5 -hours) at an extremely high temperature (1300-1350°C [in furnaces equipped with poly-crystalline-silicon or silicon-carbide tubes], are now considered standard annealing conditions) in N_2 (with about 2% oxygen added, to protect the silicon from the pitting phenomenon which occurs when silicon is annealed in pure-nitrogen at high temperatures), or Ar. It is generally agreed that the higher the annealing temperature, the better the quality of the SIMOX material. An anneal in pure-nitrogen can also be performed if a cap-layer of oxide (usually deposited by CVD) is formed on the wafer after the implant, and prior to the anneal.

This extremely-high-temperature annealing-step has three functions: **1)** It drives the chemical reaction which forms a continuous-buried-amorphous-oxide (BOX) layer of silicon-dioxide near the peak of the

implantation. (The choice of a sufficiently-high O^+-implant-dose and energy provide the required oxygen concentration to enable stoichiometric SiO_2 to be formed); **2**) It removes many of the defects (dislocations) formed during the ion-implantation step; and **3**) It dissolves any SiO_x-precipitates in the silicon device-surface-layer and denudes this layer of oxygen as well (see Chap. 2 for a discussion of denuded-layer formation in CZ-silicon-wafers). To allow integrated-circuit devices to be fabricated on the wafer after the formation of the BOX, a thin layer of crystalline-Si must remain above the BOX-layer. The BOX-layer depth is defined by the depth of the ion-implantation and it's thickness is limited to about 0.5-μm.

The evolution of the SIMOX structure during the high-temperature-anneal (as mentioned above) is quite complicated, but it is useful to summarize the phenomena that occur from a thermodynamic perspective. During annealing, both dissolution of the oxide-precipitates and precipitation of the dissolved-oxide takes place in the oxygen-rich silicon-layers. Oxygen diffuses out of the top silicon-surface, initially denuding the zone near the silicon-surface. Within the bulk of the silicon, small precipitates dissolve into silicon and large precipitates grow from the dissolved oxygen. At any given temperature (and for a given concentration of oxygen in the silicon), there exists a *critical precipitate radius*, r_c, below which a precipitate will disappear, and above which it is stable. The critical precipitate radius increases with temperature, and becomes essentially infinite at very-high temperatures. At temperatures above 1300°C the only stable precipitate is the buried-oxide-layer itself, which has an essentially infinite radius-of-curvature. The complete dissolution of all precipitates at annealing temperatures greater than 1300°C, and the diffusion of dissolved oxygen towards the buried-oxide-layer explain the thickness increase of the BOX when high-temperature annealing is performed.

The fact that a thin-layer of high-quality crystalline-silicon must exist above the BOX-layer implies that the lattice structure of the silicon in this region must maintain a high-degree of crystallinity in the face of the high-dose oxygen-implant step. The high oxygen-implantation-dose does amorphize the layer of silicon that lies above the projected range. However, it would be unacceptable if complete amorphization occurred, and a polycrystalline layer was formed during further annealing. This would happen if the temperature of the wafer was kept too low during the oxygen implant step. On the other hand, by keeping wafers at a temperature greater than 500°C during the implant (by a combination of

beam power and external heating), the amorphization damage anneals out during the implantation process ("self-annealing"), and a high-quality crystal-structure is maintained in the silicon layer at the wafer surface. By maintaining such an elevated temperature during the implant, the implanted oxygen also tends to diffuse away from the vicinity of the projected range to form a broad, flat concentration plateau. Note that the most commonly used implantation temperatures range between 600°C and 650°C. This is because SiO_2 precipitates form in the silicon overlayer at higher temperatures (700-800°C), mostly near the bottom interface. However, large precipitates are also observed near the surface region of the silicon film (and this therefore sets an upper limit of 700°C on the implantation temperature).

3. If the device requires thicker silicon layer than can be achieved with the depth of ion-implanted oxygen, an epitaxial-Si layer can be deposited to provide this additional thickness.

The are several shortcomings associated with the high-dose SIMOX technology, including the following:

1. The formation of SIMOX wafers requires extremely high doses of oxygen, and this necessitates very-long implantation times (which, as we shall see, contributes to the high-cost of SIMOX-wafers). The need to anneal the wafers at the extreme temperatures exceeding 1300°C may also be a limitation for 300-mm wafers.

2. Transition metals may be sputtered from the beam-line or end-station components during the implantation. These metals can deposit onto the SIMOX wafer surface and can then be driven into the silicon by the wafer-heating and subsequent thermal processing. Their presence in the silicon will enhance junction-leakage and reduce the minority-carrier-lifetime of the devices. Metallic contamination is present in all implant processes, but since high-dose SIMOX implants are nearly a thousand times greater than those of conventional high-dose source/drain implants, the metal contamination problem is potentially increased by approximately the same degree.

This issue was especially troublesome during the early days of SIMOX, when implanters were not yet built to limit transition-metal-sputtering. Since that time, however, great efforts were made to reduce the metal contamination sources, and silicon or silicon-dioxide shields have been placed in implanters such that the beam never "sees" any metallic part. The