As MOSFETs are scaled below 100 nm, this effect becomes the dominant factor that controls the value of R_{sp}. We discuss this issue further in Sect. 5.3.5.

Since increasing the concentration of dopants in the shallow SDE-regions is a difficult task, other ways to reduce the value of the resistance of the SDE-regions have also been sought. From Eq. 5-6 it can be seen that if the length of the SDE region could be reduced, the value of its resistance would also be decreased. While this can be accomplished by decreasing the width of the sidewall spacer, this is not practical in the conventional deep-micron MOSFET-structure. That is, since a deep contact-region is needed (to accommodate the reacted salicide-layer, see Sect 5.3.1.3), the lateral-straggle of the deeper contact-implant would penetrate under the gate edge too far if such a narrower sidewall-spacer is used. However, as we will discuss in Sect. 5.3.2, incorporating elevated-source/drains into the MOSFET structure would permit the use of such narrow sidewall-spacers, thus allowing the resistance of the SDE-region to be made smaller.

5.3.1.3 Parasitic-Resistance R_{co} of the Deeper Contact-Region of the Drain:

Although the contact-region of the drain-structure has a deeper junction-depth than the SDE, it can also contribute significantly to the parasitic series resistance of the drain structure, namely through the *contact resistance*, R_{co}. That is, a salicide ($TiSi_2$ or $CoSi_2$) is formed on the surface of the contact-region by sputter-depositing a layer of Ti (or Co) onto the wafer. Then the Ti (or Co) is reacted with the silicon at elevated temperatures, forming the salicide layer. This salicide-layer (and the kinetics of its formation) play a role in the value of R_{co}.

A simple model for R_{co} (or $R_{co}W$) is given by:[21]

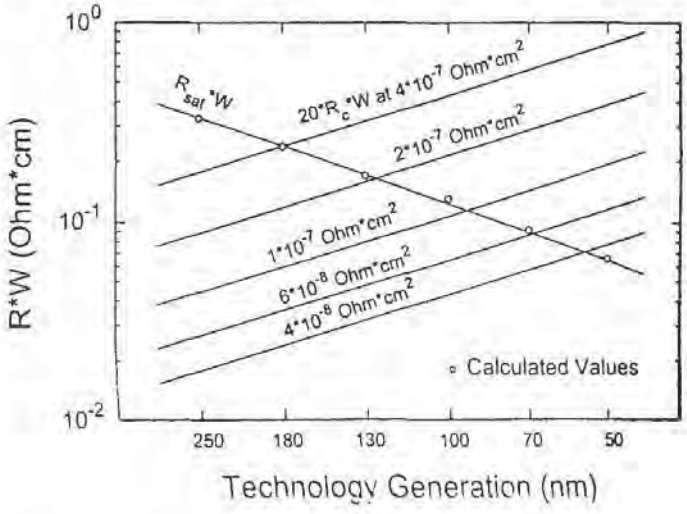


Fig. 5-18 Allowed values of drain (or source) contact-resistance.[23]

TSMC-1034
U.S. Patent 8,587,076

$$R_{co}W \geq \rho_c/2\,L_L \qquad (5.8)$$

where ρ_c is the specific contact resistivity of the silicon-to-silicide interface (in Ω-cm^2). This can again be compared with the channel-resistance R_{sat}. Figure 5-18 gives values of $20R_{co}W$ for a 10% contribution of (source + drain) contact resistance. To have a 10% or less contribution, the data in Fig. 5-18 indicates that ρ_c must be in the range of 10^{-7}-Ω-cm^2 for 0.25-μm MOSFETs, but must get down to around $4x10^{-8}$-Ω-cm^2 for 50-nm MOSFETs. It has been demonstrated that for a barrier height of about 0.5-eV, the required doping-density needed to achieve such a low value of ρ_c is above 10^{20}/cm^3. This means that the silicon regions in contact with the silicide will have be very-heavily doped, a condition that may not be possible to attain in 50-nm MOSFETs with conventional structures, for reasons we describe next.

When the silicide is formed on the source/drain contact-regions by the reaction of Ti (or Co) and silicon, some of the silicon in the substrate is consumed (i.e., when $TiSi_2$ is formed, the silicon consumed is about 75% of that of the thickness of the final silicide film, while for $CoSi_2$, even more silicon is consumed – about equal to the thickness of the $CoSi_2$ that is formed). This has two repercussions. First, the contact-region of the drain must deep enough so that the junction is not entirely lost as a result of the silicide-formation. To keep the implanted dopants from negating the benefits of the shallow SDE-junction (as a result of lateral-straggle that causes the implanted species to penetrate laterally under the implant-mask edge) a sufficiently wide gate-sidewall spacer must be used to mask the contact-region-implant. This impacts the resistance of the SDE-region, through an increase in l_{SDE}. Second, the silicon consumed is that nearest the surface, just where its doping concentration is the highest. The interface between the resultant silicide and the silicon is far below the initial silicon surface, where the implanted doping-concentration is considerably lower (which will impact the value of R_{co}). Furthermore, the dopant in the silicon (especially boron) is known to segregate into the silicide layer, further depleting the doping-concentration at the silicide-silicon interface. Since the contact-region of the drain cannot be made arbitrarily deep (due to the fact that lateral-straggle also increases as the projected-range of the implant is increased), it is not clear that it will be possible to make contact-regions with doping concentrations that are adequately high to achieve the required low-values of ρ_c. The use of elevated-source/drain regions (Sect. 5.3.2) would also allow a way around this obstacle.

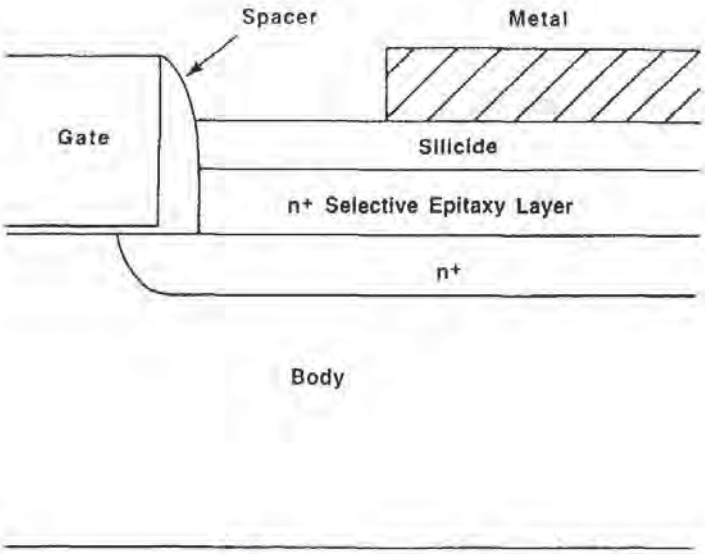


Fig. 5-19 Schematic of elevated-source/drain contacting structure.

5.3.2 MOSFETs with Elevated-Source/Drains

One proposed method for reducing the series-resistance of the source/drain structures in deep-submicron MOSFETs is to modify the basic device structure by adding so-called *elevated* (or *raised*) *source/drains* (Fig. 5-19). In this method, the heavily-doped contact is elevated above the original substrate surface by forming a selectively-grown epitaxial-Si (or perhaps Si-Ge) layer.

The MOSFET with elevated-source/drains has several advantages for devices when they are scaled to 50-nm and beyond. First, the width of the sidewall-spacer can be made much smaller. This can greatly reduce the resistance of the SDE region by decreasing its length l_{SDE}. Second, the n^+ contact layer can be sufficiently thick to contain the salicide layer without junction-leakage problems. Third, the selective-epitaxy-layer can potentially be heavily-doped during growth to obtain a lower contact-resistance. However, two selective-epitaxy processes will be needed since both n^+ and p^+ contacts must be made. The Si-epitaxial-growth steps occur after the SDE-implant. To keep the dopant of the SDE-region from diffusing too-far under the gate-edge, the selective-epitaxial-processes will have to be carried out as sufficiently low temperatures (i.e., probably below 850°C). A technique involving such a low-temperature epitaxial-process for forming elevated-source/drains was described by Miyano *et al.*[24]

In late-2001, Intel announced that the deep-submicron-MOSFET structure it intends to use for the next decade is designed with elevated-source/drains (and an SOI-structure that also employs a high-*k* dielectric [ZrO_2], SSR-channel-doping, and pocket-implants for punchthrough-control).[58]

5.3.3 Shallow-Junction-Formation for Deep-Submicron Source/Drains

The main thrust for achieving shallow SDE region junction depths is to decrease the energy of the SDE-implant. For the source/drain-structures of the 0.13-μm and 0.1-μm technology nodes, single-digit keV energies are expected to be needed for the implants that will form the drain contact regions, and sub-keV energies will be needed for the SDE-region-implants.

5.3.3.1 Ultra-Low-Energy Implants: There are three reasons that reducing the energy of the implant leads to a decrease in the junction-depth. First, the projected range R_p of the implant becomes smaller as the energy is reduced. Second, the straggle and ion-channeling effects both scale with R_p (which is especially important for B implants). Third, TED that occurs during the implant-annealing step is reduced (see Sect. 5.3.3.2). It turns out that for sub-keV B-implants the channeling-effects become relatively unimportant (compared to the broadening of the implant doping profile caused by TED during the annealing step). For this reason, elaborate measures to reduce channeling (such as performing a pre-amorphization-implant prior to the low-energy SDE-implant) are not likely to be used for MOSFETs with gate lengths of 0.1-μm and smaller.

The difficulty with implementing such low-energy-implant processes in manufacturing applications is achieving the productivity that the industry has been accustomed to getting at higher-implant-energies. For example, at energies greater than 10-keV, a high-current-implanter with a beam-current between 10- and 25-mA can implant 230 wafers per hour with the high dose of $4x10^{15}/cm^2$. (Note that beyond 0.18-μm, the doses of the SDE will be in the $1\text{-}2x10^{15}/cm^2$ range, and the doses of the contact-regions of source/drains will be even higher.) At sub-keV energies, the maximum beam-current available on commercial implanters in the mid-1990s timeframe was in the 1-2-mA range. Even after a reduction in dose (made possible by the decreasing energy), the productivity is only fraction of what it previously was. The decreased beam-current at lower energies follows from Child's law, which states that the maximum ion-current that can be extracted from any cathode, drops off as $V^{3/2}$ where V is the extraction voltage. While this basic constraint can be somewhat circumvented (by extracting a larger current at a higher energy and then decelerating the beam

to a final lower energy; i.e.., what implant vendors refer to as "decel"), the effective beam-current still decreases with the final energy. With decel, care must also be taken to minimize energy-contamination caused by ion-neutralization before (or in) the decelerating field. That is, ions neutralized while being transported in the beamline will not be counted as part of the dose. Furthermore, they will be implanted at much higher energy, possibly causing deeper junctions (thus defeating the central purpose of using an ultra-low-energy implant).

There are two commercially available high-current, low-energy-implanter configurations: **1)** *batch* (or *multiwafer*) implanters; and **2)** *single-wafer* implanters. Batch type implanters are sold by Axcelis (GSD/200E® model) and Applied Materials (Quantum LEAP®). A single-wafer, high-current, low-energy-implanter (with a scanned ribbon beam) is offered by Varian Semiconductor (VIISta-80®). The vendors of the "batch" type low-energy-implanters argue that the throughput is significantly higher than for single-wafer machines. The batch systems use a fixed beam, while the single-wafer systems use a scanned beam. The tuning time for a fixed beam is always shorter than that needed for a scanned beam, Hence, the set-up time for a batch system is smaller. Capacity modeling studies have shown that fewer batch low-energy-implant-systems are required (e.g., 3 to 4 for a fab that starts 5000-wafers per week with a 0.18-μm process, versus 5 to 9 single-wafer-implanters), primarily because of the fast tuning in the batch machines. This reduces the overall capital-cost and cost-per-wafer.

The advantages claimed for the low-energy, single-wafer-implanters are the following: **1)** the angle of the implant with respect to the wafer surface can be controlled more accurately in a single-wafer-implanter; **2)** sputtering of metals from the spinning disk (the largest source of contamination in batch-implanters) is eliminated; and **3)** a 70°-angle analyzing-magnet allows the VIISta-80® to essentially eliminate the problem of energy contamination.

While a significant effort was being expended to improve the productivity by increasing the beam current,[25,26] the basic capability of being able to implant at an ultra-low-energy (i.e., sub-keV) already existed by 2000. (That is, by 2000 implanters were available that could provide 4-mA at 0.2-keV.) This allowed researchers to investigate the limits of shallow SDE-junction-depths with sub-keV implants and conventional RTA methods. Drain extensions with junction-depths that meet 0.1-μm technology requirements were demonstrated.[27] It appears, however, that it may not be possible to improve on these results unless either the extension-resistivity-specification is significantly relaxed or new annealing technologies are developed which provide a smaller thermal budget

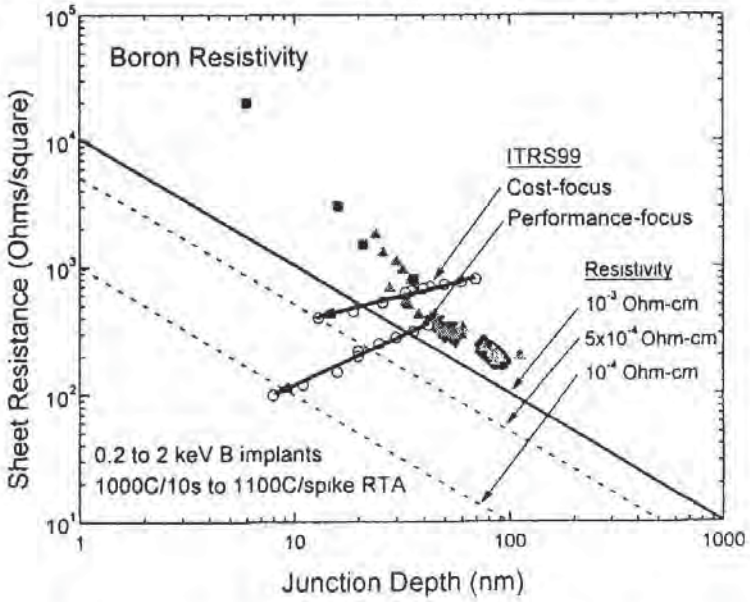


Fig. 5-20 Sheet resistance and source/drain junction depth expectations for CMOS transistors according to the ITRS99 study for cost and performance-focused product and gate sizes from 180 nm towards deeply-scaled devices with gates ~20 nm. Also shown are experimental results for sub-keV B^+ implants with "spike" RTA in CZ-Si materials.[59]

and are compatible with CMOS processing. That is, in order to meet the decreased resistivity requirements of the scaled MOSFETs, it is necessary to implant a larger dose or increase the annealing temperature to activate more of the smaller dose implant. The issue is illustrated in Fig. 5-20, where the ITRS99 expectations for the source/drain-extension junction-depth and sheet-resistance are compared to experimental results for boron-doping with sub-keV B-implants and a "spike"-RTA process (with a high-temperature thermal-cycle of 1050°C for 1 sec). All of the experimental results are above the ~10^{-3}-Ω-cm resistivity "wall" (the dark downward-sloping line in this figure). The shaded squares and triangles in Fig. 5-20 are actual experimental data, whereas the open circles are expected results that have not yet been attained. The resistivity requirements for "high performance" MOSFETs are ~10^{-4}-Ω-cm for 20-nm MOSFETs.

As compared to the SDE-regions of the source/drain, higher doping-concentrations are needed for the contact source/drain-regions. This is because these regions are used to contact the silicide-layer, and high doping-concentrations help to reduce the value of R_{co}. Since these regions are not adjacent to the channel, they do not have as great an impact on short-channel behavior as the SDE-junctions. Consequently, doses and implant-energies can be higher than those permissible for SDE-regions, even when B-implants are used.

By 2001 the trend of using lower implant energies for the SDE-implant for each generation appears to have reached a practical limit in terms of device performance (especially for *p*-channel transistors). That is, 0.5-keV boron implants appear to be the lower useful limit for shallow junctions, not just for 0.13-μm but for all technologies.[28] Below 0.5-keV, the implant becomes less productive, more costly, and the final junction performance (i.e., the simultaneous achievement of junction-depth and resistivity) is the same. The performance limit with a 0.5-keV boron-implant and spike-anneal is a junction-depth of about 35-nm and a sheet resistance of about 350-Ω/sq. It appears that to further reduce either value, the other must be increased.

5.3.3.2 Transient-Enhanced Diffusion: The second key reason for developing dopant implant processes with energies in the sub-keV range (besides reducing the as-implanted depth), is to minimize the transient-enhanced diffusion (TED) of the dopant. When high-energy ions enter the silicon lattice, they displace silicon atoms from their lattice sites, creating pairs of vacancies and interstitials. When the post-implant anneal is initiated, most of these interstitials and vacancies recombine, but some interstitials remain behind (in an amount about equal to the quantity of implanted atoms – most of which now occupy the sites that were formerly filled by the still-displaced silicon interstitials). The fact that each implanted atom produces about one residual interstitial after a very short anneal is called the "*+1 approximation.*" TED can be explained by relating its enhanced-diffusion effects to the presence of these excess interstitials for a wide range of low- and medium-dose implants, and low- and medium-energies.

The residual excess interstitials rapidly coalesce into such unstable extended-defects as {311}'s and also into more-stable extended-defects, such as dislocation loops. While this coalescence occurs because there is a lower system-energy when the interstitials populate extended-defects (compared to when they exist as individual interstitials), the defects are still metastable. Thus, they dissolve with continued annealing (especially the {311}-defects). During their dissolution they emit interstitials into the silicon. The rate at which these defects dissolve depends on the activation-energy and annealing-temperature. Nevertheless, as long as these defects exist, an interstitial supersaturation is maintained (i.e., a higher concentration of interstitials is present than would exist in the absence of the extended defects). Since boron diffuses by an interstitial mechanism, its diffusivity is enhanced by the excess interstitials - with the time-averaged diffusivity-enhancement equal to the time-averaged interstitial-supersaturation. In any case, both the interstitial-supersaturation and the

diffusivity-enhancement soon end after the defects have dissolved. The enhancement is thus transient.

The *increase* in depth of a shallow SDE-junction ΔX_j due to TED can be expressed in terms of the initial-supersaturation of interstitials and the time of the interstitial defect dissolution as:

$$\Delta X_j^2 \propto N R_p \exp [- (-1.4 \text{ eV})/kT] \quad \textbf{(5.9)}$$

where N is number of interstitials initially trapped in the defects (approximately equal to the implanted dose, as noted above) and R_p is the projected ion range (which is where the peak concentration of the excess interstitials is initially located). The linear dependence of N simply represents the fact that the larger the dose, the larger the supersaturation, and hence the greater the diffusivity enhancement. The linear dependence on R_p has been demonstrated experimentally. The nearer the excess interstitials are to the surface, the faster they reach the surface (where they are annihilated). The linear dependence on R_p explains why use of ultra-low-energy implants reduces TED effects. That is, if an ultra-low-energy implant is used, R_p will be small. Thus, the location of the interstitials will be closer to the surface, allowing them to escape there faster. Thus, the duration of the transient is shortened, making the value of ΔX_j smaller.[29] (Later work indicated that for implants below 1-keV in energy, TED is even further reduced because there is less lattice damage per ion.[30])

Another important dependence contained in Eq. 5-9 is the *negative* activation-energy of ΔX_j^2, which arises from the fact that the interstitial-supersaturation due to the presence of the extended-defects is larger at a lower temperature. The negative-activation-energy implies that if the defects are annealed out at a low temperature, the final junction will be deeper than if the defects are annealed out at a higher temperature. This is one of the main reasons why the S/D anneals are currently carried out in an RTA tool rather than in a conventional batch furnace (that has a ramp-up rate of a few degrees per minute). In an RTA chamber, significantly less time is spent during the temperature ramp-up at a lower temperature (at which the diffusivity-enhancement is larger).

5.3.3.3 Rapid-Thermal-Annealing of Shallow Junctions: To form the shallowest junction possible using ultra-low-energy implantation, it is necessary to minimize the thermal-budget after implant. (Thus, as noted above, annealing such implants is carried out in RTA tools to realize good-quality shallow junctions.) In the ideal case, the as-implanted-dose would spend "zero" time at the temperature that would electrically-activate the dopants and heal the lattice damage from the implant. (Exposure to a high-temperature for a very-short time is effective

because the activation-energy for dopant-activation is typically higher than that for diffusion, thus electrical activation is favored while limiting diffusion. Reported values for boron are 5.4-eV and 3.4-eV for activation and diffusion, respectively.) Attempts to reach this ideal annealing condition have led to the development of so-called "spike" anneals, which consist of very fast ramp-up and ramp-down of temperature, with virtually zero-time at the peak temperature. Experiments have shown that the difference in the percentage-of-activation and sheet-resistance between spike-anneals and those that keep the wafer at temperature for 10-sec is negligible. On the other hand, for the 10-sec-at-temperature anneals, the junction-depth increases with temperature.[88] This indicates that spike anneals are more effective at higher peak-temperatures.

In practice, spike anneals have an effectively finite-time at temperature. Also, in reality, while ramp-up rates as fast as about 200°C/sec are possible, the ramp-down is limited to approximately 50-80°C/sec. As a result of these deviations from ideality, it is useful to determine what is a sufficiently-fast ramp-up rate. The negative-activation-energy of TED seems to indicate that as high a ramp-rate that can be attained, should be used during a spike anneal. However, experiments show that for such ultra-fast ramp-rates, the advantage gained by going from moderately fast (e.g., 100°C/sec) to ultra-fast ramp-up rates (200-400°C/sec) is relatively small for ultra-low-energy implants. Such a small advantage must be balanced against the loss of process control (e.g., across-wafer thermal-uniformity and temperature-repeatability) which inevitably occurs when such ultra-fast ramp-up-rates are used. The reason the advantage is not very great is that TED is already sufficiently minimized at 100°C/sec.[32] This reduction in thermal-budget during ramp-up is negligible compared to the portion of the thermal budget which cannot be reduced (the finite-dwell-time at temperature and the relatively low ramp-down-rate).

The data shows that the largest reduction in junction-depth with increased ramp-up-rates occur for the lowest-energy-implants (see Fig. 5-21).[31] Since TED decreases with decreasing implant-energy, it can be inferred that the advantage is primarily due to the overall thermal-budget and not just of the TED. It is therefore evident that without the ability to increase the ramp-down-rate (when it is relatively smaller than the ramp-up-rate), the advantage from increasing the ramp-up-rate will quickly saturate at ramp-up-rates greater than the ramp-down-rate. In addition, it has been determined that the ramp-rate is not as good a measure of the quality of a spike-anneal as is the time spent diffusing the dopant. That is, since about 90% of the dopant-activation occurs in the short-duration from approximately 50°C below the peak-temperature to the peak-temperature

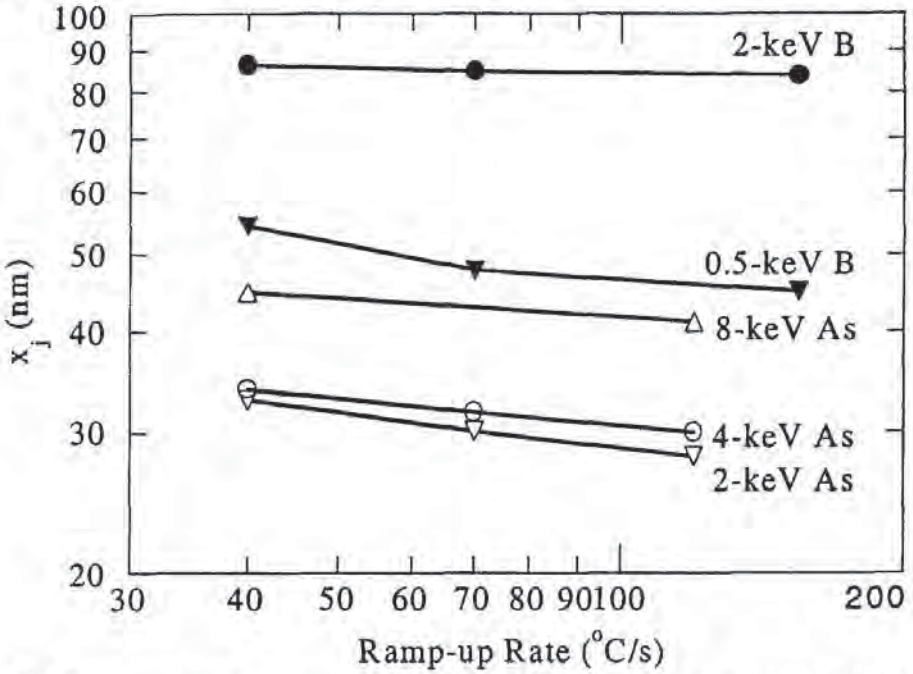


Fig. 5-21 Junction-depth, measured at a concentration of $1x10^{18}/cm^3$ after spike-annealing at various ramp-up-rates. The ramp-down-rate was the same in all cases. For both As and B the advantage from a faster ramp-up is largest for the shallowest implant. This implies that the reduction in TED (known to decrease with implant-depth) is not the dominant mechanism responsible for the larger reduction in junction-depths observed for the shallower-implants with increasing ramp-up-rates.[31]

and back to 50°C below, this interval of time is really a better specification of the quality of a spike-anneal.[28]

Figure 5-22 is a survey of published spike-anneal data that was presented in Ref. 32. It can be seen that many of the data points are coincident, despite having been taken from different implant-and-annealing conditions (energy, dose, and ramp-rate). It is evident that more than one set of process conditions can be used to arrive at the same junction properties.

A variety of RTA-tool configurations are commercially available, including the following: **a**) linear lamps, double-sided heating; **b**) axisymmetric lamps, single-sided heating; **c**) arc lamp; **d**) hot-plate; and **e**) hot-wall RTP. Several general references are available that give detailed descriptions of these types of RTP tools.[33,34,35] Data on shallow junctions formed with the various RTP-tool types can be found in the following references: lamp-based RTP systems (36); hot-plate (susceptor – 37); and hot-wall RTP (38).

New ways to increase the value of the electrically-active-dopant in shallow-junction regions are also being actively being researched. Some of these novel techniques include: **1**) laser anneals;[39] **2**) strain-compensation implants

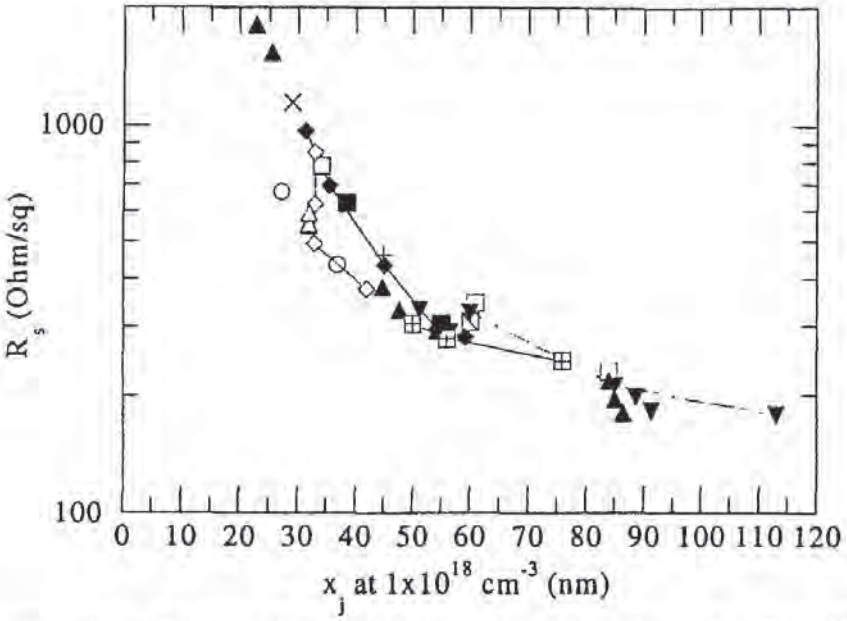


Fig. 5-22 Comparison of *p*-type sheet-resistance and junction-depth data for spike-anneals using various implant-and-annealing conditions. The data set includes a 400°C/sec ramp-up-rate and 0.2-keV implant-energy. Note that all the data points fall in a rather narrow band. No one has been able to make a 30-nm junction with a 300-Ω/sq sheet-resistance using ion-implantation and conventional rapid-thermal-annealing.[32]

(e.g., by adding Ge-atoms to the silicon lattice, their larger size compensates for strain caused by the smaller B atoms. This increases the solubility of the boron atoms in the silicon.);[40] and **3**) using RTP-tools which have a short lamp-response-time, which increases the ramp-down-rate.[41]

5.3.3.4 Dopant Loss for sub-keV Implants from Self-Sputtering, Out Diffusion, and Surface Oxide: Several factors associated with ultra-low-energy implants give rise to loss of some of the species that are implanted, and they are discussed in this section. The effects that lead to such dopant loss are: self-sputtering, out-diffusion, and the presence of a surface oxide during the implant process. When the dose-loss from all three of these effects are combined, the loss can be large enough to pose a significant problem for ion-implant processes carried out at energies of 0.5-keV or lower.

The first phenomenon associated with dopant loss in ultra-low-energy implant processes is *self-sputtering*. That is, when any solid-surface (the target) is struck by an ion-beam, sputtering of atoms from the target will occur. The number of atoms sputtered for each bombarding ion rises as the energy of the ion decreases. In particular, when the ions being implanted have energies in the 0.2-1-keV range, significant sputtering of the target will occur. For example, when 0.5-

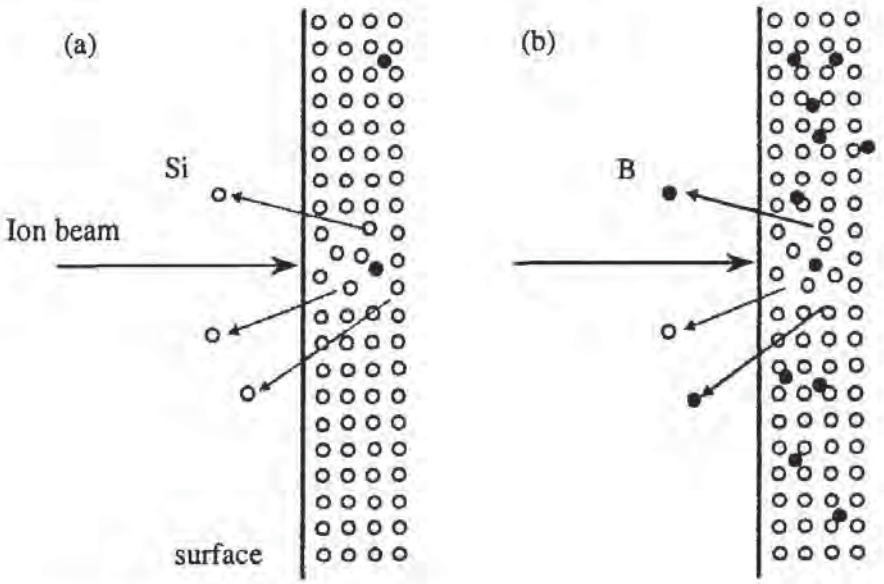


Fig. 5-23 The dose-loss due to self-sputtering depends on how much dopant has already been implanted. (a) At the start of the implant most of the sputtered atoms are silicon. (b) As the composition of the near surface layer becomes altered, with an increasing fraction of them being boron, more boron atoms are sputtered.[59]

keV B-ions strike a silicon target, about one Si-atom is sputtered for every four implanted B-atoms. Since at such low-energies the implanted species also come to rest near the Si surface, some of these implanted B-atoms themselves will be subject to sputtering. In fact, it is expected that with increasing B-dose, more and more of the implanted B-atoms will be sputtered from the surface. (This phenomenon is depicted in Fig. 5-23.) Since additional B-implanted-atoms will be sputtered with increasing dose, this implies that the fraction of implanted dose that is retained in the silicon wafer will decrease with increasing dose. Given the difficulty associated with achieving a large beam-current at such low-energies, a further reduction in the retained dose is very undesirable. That is, while the dopant-loss could be compensated by increasing the implant dose, this will further decrease throughput (which is already too low).

Figure 5-24 plots the retained-dose compared with the nominal-dose for various energies.[32] It indicates that at 2-keV essentially the entire dose is retained. But at implant-energies of 0.5-keV and 0.2-keV, a significant fraction of the implanted-dose has been sputtered out, with the missing fraction rising with increasing dose. For a nominal-dose of $1x10^{15}/cm^2$, sputtering causes a 10% reduction in the retained dose at 0.5-keV, and a reduction of 20% of the retained dose at 0.2-keV.

If BF_2^+ is used to implant B at low-energies, the sputtering problem is expected to be even worse than if B is implanted. That is, in BF_2^+-implantation, for each implanted B-atom that contributes to the retained dose, there are now 3 atoms (one B and 2 F) which can cause sputtering (compared to just one B-atom for an implant process that uses B^+-ions instead). For example, for a nominal dose of $1x10^{15}/cm^2$, we stated that a 0.5-keV B^+-implant will cause a 10% reduction in retained dose. However, if a BF_2^+ implant is carried out so that the B-atom in the BF_2^+-ion has the same energy as in the 0.5-keV B^+-process (i.e., the BF_2^+-ions have energies of 2.2-keV), the reduction in retained-dose increases to ~20%. The above circumstances imply that caution must be exercised when comparing junction data from the literature for different sub-keV implants or for different species. A correction needs to be applied for the different retained-doses to allow meaningful comparisons to be made.

The second effect that gives rise to dose-loss is *out-diffusion.* Out-diffusion of implanted species occurs during the annealing steps that are performed after the implant. As the projected-range R_p of the dopant is decreased (as a result of implanting at lower energies), the out-diffusion grows worse. The reduction in retained-dose after a 1000°C/15-sec-anneal is approximately 10% for an implant energy of 0.5-keV, but increases to 15% for 0.2-keV energies.

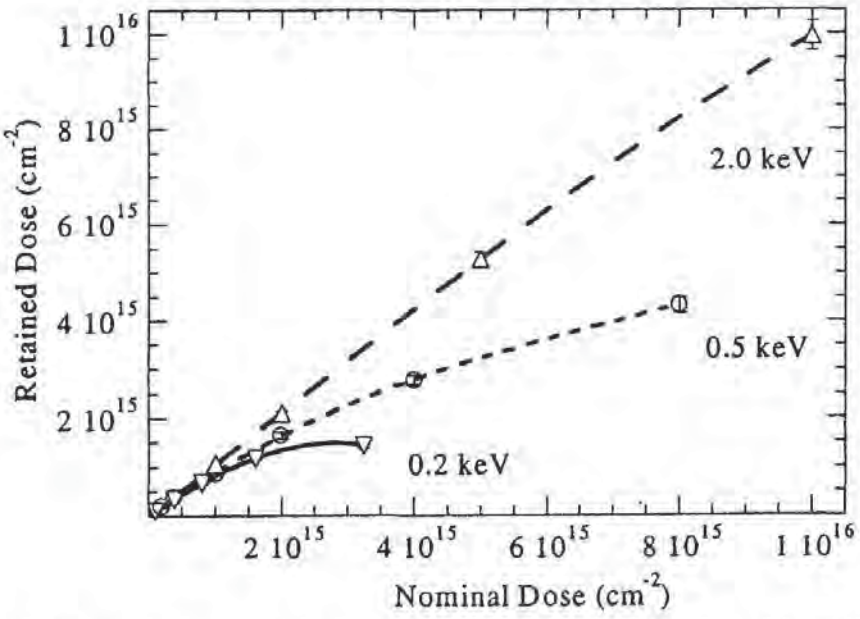


Fig. 5-24 Retained B-dose as a function of nominal B-dose implanted into (100)-silicon wafers with an as-received native oxide. The trend lines are 3rd order polynomial spline-fits. For a nominal B-dose of $1x10^{15}/cm^2$, sputtering leads to a reduction of ~10% in the retained dose at 0.5-keV, and a reduction of ~20% at 0.2-keV.[52]

The final cause of dopant-loss in sub-keV implants is due to the presence of an *oxide on the silicon-surface during implant.* Such an oxide is often deliberately grown to protect the silicon surface from metal contamination during the implant process (see Chap. 10, Vol. 1, 2nd Ed.). While this protective oxide layer can be made quite thin, it nevertheless has a finite thickness. It is clear that as the implant energy is reduced, a larger fraction of the as-implanted dose will be located in the protective oxide. Since the oxide is stripped off at the end of the implant step, the dopant that remains in the oxide cannot contribute to the formation of the junction. Approximately 25% of the B implanted at 0.5-keV is lost by removal of a 15-nm-thick surface-oxide, while more than 40% is lost for 0.2-keV implants. With these results, it is possible to estimate the equivalent-dose needed at lower energies (such as 0.2-keV) to make up for the larger loss as compared to 0.5-keV implants. For example, it has been calculated that to successfully implant a dose of $4x10^{14}/cm^2$ B-atoms at 0.2-keV into the silicon beyond the protective oxide, the beam-current would have to be as much as 60% greater than the beam-current at 0.5-keV to obtain an equivalent throughput process.[32] To circumvent the oxide dose-loss problem, some ultra-low-energy implant processes are being carried out without the use of a protective oxide.

In summary, the dose-loss effects are significant for sub-keV implant processes, and the loss increases as the implant-energy is reduced. For the extreme case of 0.2-keV implants, as much as 40% of the B-dose can be lost due to the combined effects of self-sputtering, out-diffusion, and native-oxide removal for the high doses needed for an SDE-implant.

5.3.4 Effect of the Overlap of the SDE-Region and Gate-Edge on MOSFET Performance

In addition to being able to carefully set the junction-depth X_j and length of the SDE region l_{SDE}, it has also been reported that the amount of overlap of the SDE-region with the gate-edge is also critical for optimum device performance.[42,57] That is, if the overlap of the SDE-region and the gate-edge is too great, the deep-submicron MOSFET will suffer excessive leakage (due to larger short-channel effects that arise if the channel-length is too small). Excessive leakage means that the off-state current exceeds its maximum design value (e.g., 1 nA/μm). In addition, the gate-to-(drain/source) overlap-capacitance increases, which negatively impacts device switching speed.

On the other hand, if the SDE-to-gate-edge-overlap is too small, the ON-state saturated-drain-current I_{Dsat} will decrease because the SDE-region and the end of the channel are not electrically well-coupled. Thus, an excessively-high-resist-

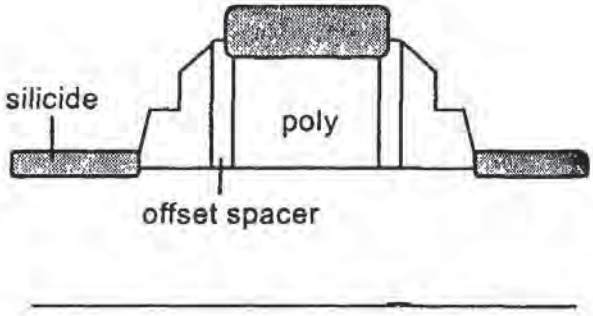


Fig. 5-25 Schematic of a deep-submicron MOSFET with an offset-sidewall spacer.[57]

ance-region separates them (and this increases the parasitic resistance of the source/drain regions, which leads to a decrease in the value of I_{Dsat}).

Controlling the position of the edge of the SDE region with respect to the gate edge can be done in the following way. First, the implant-and-annealing processes for forming the SDE (and contact) region of the drain structure are determined, and the drain structure is formed using these process conditions. However, an offset-spacer is fabricated on the side of the gate after the gate-poly is etched (but before the SDE implant step is carried out, Fig. 5-25).[57] The width of this off-set spacer is varied from zero (no-spacer) to some upper-bound (e.g., 50-nm). Measurements are then made for the MOSFET having different offset-spacer-widths. The offset-spacer-width that yields optimum device characteristics (i.e., the best combination of off-current and I_{Dsat}) is taken to be the value that produces the correct overlap of the SDE-region and the gate edge.

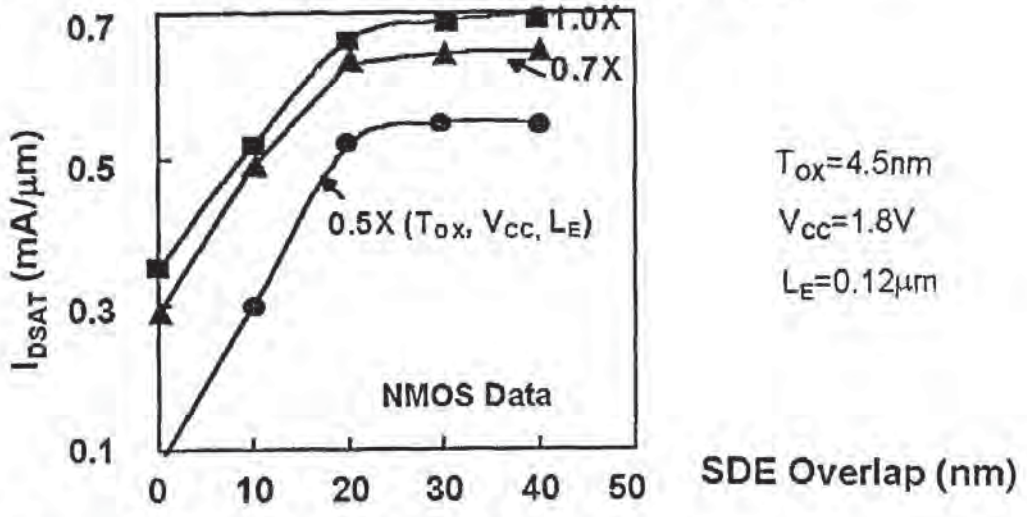


Fig. 5-26 I_{Dsat} versus SDE overlap (NMOS).[42]

The results of such a study are shown in Fig. 5-26. In this case, the effects of SDE-overlap on 0.25-μm NMOSFETs were analyzed. The arsenic-implanted SDE junction-depth X_j was measured to be 50-nm. Since the lateral diffusion of the SDE junction under the gate-edge is estimated to be 0.6-0.7 times the vertical depth (about 30-nm), minus the width of the offset spacer, the SDE-overlap for each offset spacer width can be determined. Fig. 5-26 shows that if the SDE-overlap is less than 20-nm, the value of I_{Dsat} falls off rapidly. Thus, a minimum overlap of ~20-nm is recommended by this study. Information as to how to form the offset spacer was not given, but a method (such as re-oxidizing the poly-silicon gate-structure after it has been etched) is one possible way of doing so.

5.3.5 SDE-Junction Abruptness

The SDE-region must not only have the correct junction-depth, average-doping-concentration, and position under the edge of the gate, but recent studies have shown that there must also be a highly-abrupt lateral-doping-gradient at the interface of the SDE and the MOSFET channel-regions.[43,44] Not only does such an abrupt-doping-gradient help limit the lateral-encroachment of the extension-depletion-regions into the channel, but more importantly it minimizes R_{sp}. The design rules for lateral-junction-abruptness are that the doping-falloff must be at least one decade in a distance equal to ~3% of the size of the gate-length (for a spreading-resistance less than 5% of R_{sat}). For 100-nm MOSFETs this means that the junction-abruptness must be ~3 nm/decade, which is about the profile slope of as-implanted B at 0.2-keV.[45] Since the abruptness will have to increase even further as smaller MOSFETs are designed, this implies that new techniques will have to be developed that result in final junction-doping-profiles that are even steeper than the as-implanted profiles.

The design of the implanter can play a role in the lateral-abruptness of the SDE as-implanted doping profile. That is, unlike mid- to high-energy ion-implant applications, channeling-effects are minimal for low-energy, high-current implants (because of increased channeling acceptance angle). Nevertheless, there is still a significant difference between single-wafer and batch-implanters due to implant shadowing effects. Ideally, a 0°-implant will give the most-abrupt lateral-junction because any tilt-angle larger than 0° induces shadowed areas at the vicinity of the gate pattern (leading to laterally-graded doping-profiles). If the incident beam has an angular-spread deviating from the ideal parallel beam, each of the ions having different incidence angles will make different shadows. The result is a loss of dose- and junction-abruptness at the outskirts of the gate. A single-wafer, parallel-beam-implanter minimizes the incident-angle-variations.

Spinning-disk batch-implanters use a stationary-spot ion-beam. The wafers pass through the beam due to the rotation of the disk. Such batch-implanters use a concave-shaped disk to secure wafers in their respective sites during the high-speed motion of the disk. Centrifugal force exerted on the disk surface holds them in place. Since the conical disk has a curved surface and the wafer-surface is flat, this leads to an incident-beam angle-variation across the wafer during the implant process. Depending on the dimensions of the disk, cone-angle, and wafers, this can amount to 1.5°-variation across a wafer.[46]

Recently, the abruptness of the SDE-lateral-doping was measured using both scanning capacitance microscopy (SCM)[47] and scanning tunneling microscopy (STM).[48] The latter measurement detected a tail-region of dopant at the end of the doping profile, spreading from the extension-region in the horizontal direction near the gate oxide. The cause of this anomalous diffusion tail was not understood.

5.4 ANTI-PUNCHTHROUGH STRUCTURES IN DEEP-SUBMICRON CMOS

Lateral tailoring of the channel doping profile can provide some additional flexibility to control V_T and punchthrough. Recently, such lateral tailoring of the channel profile has been carried out with the use of *large-angle-tilt implanted punchthrough stopper* (LATIPS), which are also referred to as *pocket-implants* (and are depicted in Fig. 5-27). Such a pocket-implant puts additional dopants in the channel close to the source/drain-junctions for punchthrough control, and leaves much of the channel region with a lighter doping, as needed for a low V_T.* The peak doping values of the pocket-implants must be somewhat larger than the doping-concentration values needed for a punchthrough-control layer that is uniform in the lateral channel direction between the source and drain.

These punchthrough-suppression pocket-implants are done after the source/drain-extension-implant (and before the formation of the spacer used to define the edge of the deeper contact-source/drain-implant). This ensures that the pocket-implant-dopant is self-aligned to the gate. The dopant is placed just below the active channel, adjacent to the source and drain, and is of the same dopant-type as the well. Precise control of the placement and dose of this implant within the transistor-structure is needed to obtain the intended enhancements. As noted

* As pointed out in Chap. 1, if an implant with a beam directed normally to the wafer surface is used to introduce dopants to create a punchthrough-control structure, instead of one whose direction is tilted at a large angle to the wafer surface, the resulting punchthrough-control structure is termed a *halo* structure. Such a halo structure is described in Chap. 1, and is depicted in Fig. 1-2b.

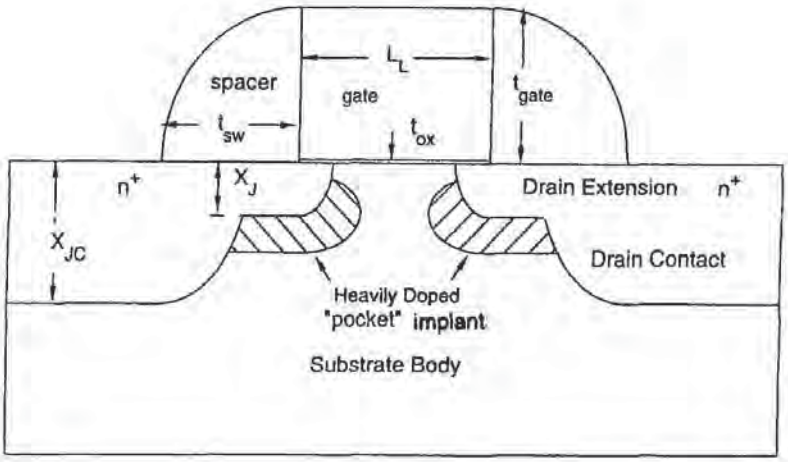


Fig. 5-27 MOSFET with heavily-doped pocket-implants for punchthrough-control.

above, the use of high wafer-tilt-angles when performing such pocket-implants offers an additional, controllable processing parameter for producing complex channel structures, without diffusion or additional masks. (The presence of the pocket-implanted-species also helps keep the source/drain-extension [SDE] junction-depths shallow, because they compensate some of the dopant in the SDE region [which, in effect, squeezes the SDE-junction closer to the surface].)

As introduced by Hori, high tilt-angles are needed for achieving greater

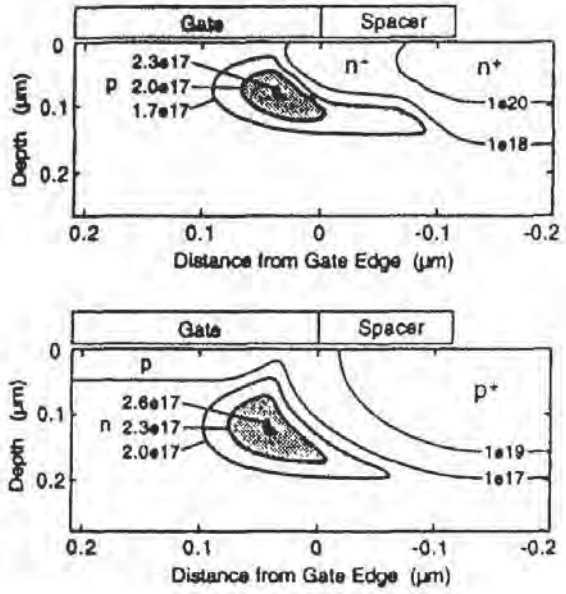


Fig. 5-28 Simulated 2-D impurity profiles of typical LATIPS structures for (a) NMOSFET with *n*-LDD region and; (b) Buried-channel PMOSFET.[49]

lateral-penetration of the pocket-implant than the adjacent source/drain-extension implant without thermal diffusion.[49] That is, the species implanted by the pocket-implant should reside outside of the final source/drain-extension-doping-profile after implant and annealing. Typically, tilt-angles are 20°-45° from the wafer normal. Figure 5-28 shows the resulting (simulated) doping profile in the channel region and in the adjacent well of a LATIPS device.[49]

To maintain device symmetry with high tilt-angle, the implants must be split into segments. Since most circuits contain transistor gates placed 90° from each other, high tilt-angle implants are typically done in four segments of equal dose ("quad-implants") with the wafer rotated 90° between segments. (As gate-widths shrink, less undercutting of the gate is required to produce the punchthrough stop effect. Consequently, tilt-angle requirements for this application are steadily decreasing below the 0.25-μm node, and are expected to reach the "low-tilt-angle" threshold of 10° by the 0.1-μmtechnology node.)

High-tilt-angle punchthrough-suppression-implants improve device performance. Junction-capacitance can be reduced by up to 65%; while the subthreshold slope of the device becomes steeper (which suppresses short-channel effects).[50] The dramatic improvement in punchthrough-resistance does not depend on extreme scaling of the S/D junction-depth or *n*-well background-doping. The reduced subthreshold-slope and junction-capacitance directly result from carefully controlled doping underneath the channel at the gate edge and the near elimination of all other undesired doping that results from less controllable doping methods.

Because of their lower-diffusivity and as-implanted-straggle, higher mass dopants can be placed more precisely in the channel. Consequently, indium (In) and antimony (Sb) are being increasingly used for this application.[51,52] Punchthrough-stop implants were the primary production application of In and Sb in ion-implantation in the 2000 timeframe. The lower susceptibility of these species to TED also helps reduce the reverse-short-channel-effect (the slight increase in threshold-voltage with decreasing effective-channel-length that occurs before punchthrough-effects dominate the smallest channel-lengths).

One proposed structure for MOSFETs with an effective gate length (L_E) of 0.25-μm is shown in Fig. 5-29.[53] This device has a very heavily doped source/drain next to a very heavily doped pocket-implant. Since the gate-length is so small, the pocket-implanted-region actually doubles for the traditional retrograde-well-implant. This aggressive device design shows the challenges of doping a 25-nm MOSFET. Very-high doping levels and very-steep doping gradients are required – with a minimum of thermal processing.

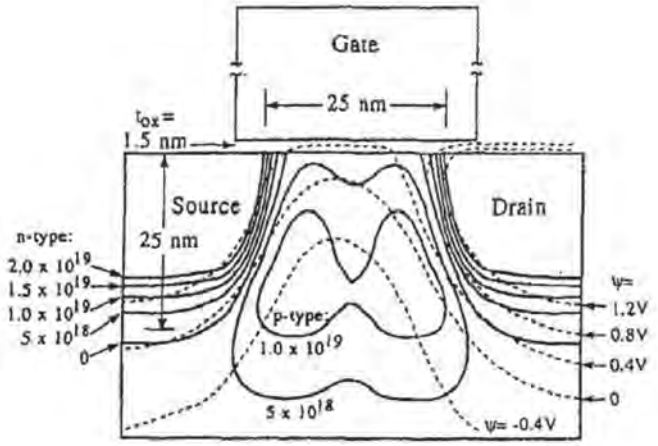


Fig. 5-29 Doping contours in 25-nm PMOSFET.[53]

An issue associated with the use of In as a pocket-implant species has recently been reported. That is, an In implant at the doses needed for the pocket-implant process causes a buried amorphous layer to form. This layer has an interface with the single-crystal silicon. The silicon at these interfaces is still crystalline, but is heavily damaged. The indium segregates to these damaged regions and becomes

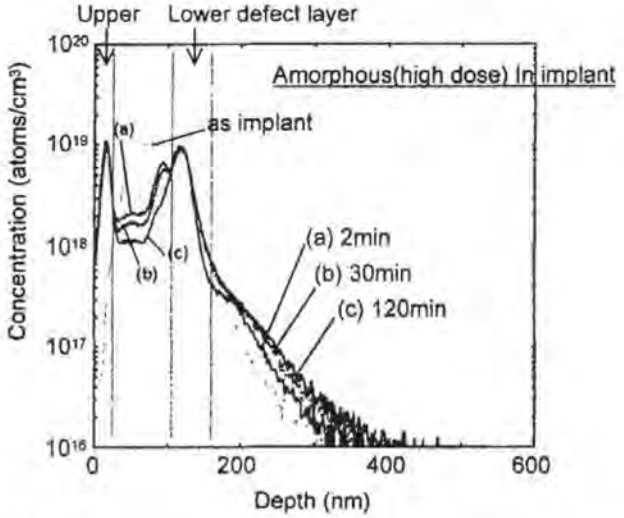


Fig. 5-30 Indium SIMS profiles for $1x10^{14}/cm^2$ implants at 200 keV annealed at 750°C for 2, 30, and 120 min in nitrogen. This figure shows large dose loss for indium. The depth of the defect layer (measured by XTEM) is shown (upper & lower defect layers).[54]

electrically de-activated (negating its ability to provide punchthrough control).[54,55] Figure 5-230 shows a SIMS profile of the In as-implanted, and after annealing. The loss of indium from its peak is seen. A double-implanted extension region (As + Sb) has been reported to be able to control the damaged region near the In pocket, mitigating the indium segregation and de-activation.[56]

REFERENCES

1. R.D. Rung, C.J. Dell'Oca, and L.G. Walker, "A Retrograde *p*-Well for High Density CMOS," *IEEE Trans. Electron Dev.*, **ED-28**, p. 1115, 1981.
2. D.M. Brown *et al.*, "Trends in Advanced CMOS Process Technology," *Proceedings of the IEEE*, December 1986, p. 1646.
3. G.J. Hu *et al.*, "Design and Fabrication of *p*-channel FET for 1-μm CMOS Technology," *IEDM Tech. Dig.* 1982, p. 710.
4. R. Martin and J.Y. Chen, "Optimized Retrograde *n*-Well for 1 μm CMOS," *Proc. Custom Integ. Circuits Conf.*, 1985, p. 199.
5. J.Y. Chen, "CMOS - The Emerging Technology," *IEEE Circuits and Devices Magazine*, March 1986, p. 16.
6. Y. Mochizuki, "Special Report: High Energy Ion Implantation Technology," Nikkei Microelectronics (Dec. 1991), p. 92.
7. K. Tsukamoto *et al.*, "High Ion Energy Implantation for ULSI: Well Engineering and Gettering," *Solid State Technology*, June 1992, p. 35.
8. R. Wijberg *et al.*, "Strong Asymmetric Doping Profiles at Mask Edges in High-Energy Ion Implantation, "*IEEE Trans. Electron Dev.*, **37-1**, January 1990, p. 79.
9. T.M. Parrill and M.J. Namaroff,: "Ion Implantation Productivity for Advanced Applications," *European Semiconductor*, July 1999, p. 44.
10. B. Adibi and V. Meniailenkov, "Transistor Performance: The Impact of Implant Doping Accuracy," *Solid State Technology*, January 2001, p. 88.
11. S. Takagi *et al.*, "On the Universality of the Inversion Layer Mobility in Silicon MOSFETs: Part I-Effects of Substrate Impurity Concentration," *IEEE Trans. Electron Dev.*, **41**, 2357 (1994).
12. P. Boullion *et al.*, "Re-Examination of Indium Implantation for a Low Power 0.1 μm Technology," *Tech. Dig. IEDM*, 1995, p. 897.
13. S.C. Williams *et al.*, "Scaling Trends for Device Performance and Reliability in Channel-Engineered NMOSFETs," *IEEE Trans. Electron Dev.*, **45**, January 1998, p. 254.
14. I. De and C.M. Osburn, "Impact of Super-Steep-Retrograde Channel Doping Profiles on the Performance of Scaled Devices," *IEEE Trans. Electron Dev.*, **46**, August 1999, p. 1711.

15. S. Venkatesan *et al.*, "Device Drive Current Degradation Observed with Retrograde Channel Profiles," *Tech. Dig. IEDM* 1995, p. 419.
16. M. Harlan and K. Petry, *Ion Implantation Technology*, 1998, June 22-26, Kyoto, Japan.
17. T. Ghani *et al.*, "100 nm Gate Length High Performance/Low Power CMOS Transistor Structure," *Tech. Dig. IEDM,* 1999, p. 415.
18. T. Mizuno, *IEEE Trans. Electron Dev.*, **47**, 756 (2000).
19. Ohguro *et al.*,"Undoped Epitaxial Si Channel in NMOSFET Grown by UHV-CVD with Preheating," *IEEE Trans. Electron Dev.,* **45**, March, 1998, p. 710.
20. International Technology Roadmap for Semiconductors, 2001 Edition, Semiconductor Industry Association, San Jose, CA.
21. K.K. Ng and W.T. Lynch,"Analysis of the Gate-Voltage-Dependent Series Resistance of MOSFETs," *IEEE Trans. Electron Dev.*, July 1986, p. 965.
22. G. Baccarani and G.A. Sai-Halasz, "Spreading Resistance in Submicron MOSFETs," *IEEE Electron Dev. Lett.,* **EDL-4**, 1983, p. 27.
23. W.C. Holton *et al.*, "Overview of Semiconductor Devices," Ch. 1 in *Handbook of Semiconductor Manufacturing Technology,* Y. Nishi and R. Doering, Eds., Marcel Dekker, New York, 2000, p. 1.
24. K. Miyano *et al.*, "Low Thermal Budget Elevated Source/Drain Technology Utilizing Novel Solid Phase Epitaxy and Selective Vapor Phase Etching," *Tech. Dig. IEDM* 2000, p. 433.
25. D.W. Duff and L.M. Rubin, "Ion Implant Equipment Challenges for 0.18 μm and Beyond," *Solid State Technology,* June 1998, p. 83.
26. T. Parrill *et al.*, "A New Era for High-Current, Low-Energy Ion Implantation," *Solid State Technoloy,* November 2000, p. 103.
27. H.J. Gossmann *et al.*, "Junctions for Deep-Sub-100-nm NMOS: How Far Will ion Implant Take Us?" *Symp. on Front End Processing* at the *Spring 2000 Meeting of the Materials Research Soc.,* April, 2000.
28. L. Peters, "Ramping the 0.13 μm Generation," *Semiconductor International*, January 2001, p. 56.
29. A. Agarwal *et al.*, "Reduction of Transient Diffusion from 1-5 keV Si^+ Ion Implantation Due to Surface Annihilation of Interstitials," *Appl. Phys. Letts.*, **71**, p. 3141 (1997).
30. A. Agarwal *et al., Mat. Res. Soc. Symp. Proc.*, **568**, p. 19 (1999).
31. A. Agarwal *et al.*, "Ultra-Shallow Junction Formation Using Ion Implantation and Rapid Thermal Annealing: Physical and Practical Limits," *Abs. Mtg. Electrochem. Soc.,* Spring 2000, Abs. 526.
32. A. Agarwal *et al.*, "Ultra-Shallow Junctions by Ion Implantation and Rapid Thermal Annealing: Spike Anneals, Ramp-Rate Effects," *Mater. Res. Soc. Symp.*, **19** p. 568 (1999).

33. P.J Timans *et al.*, "Rapid Thermal Processing," Ch. 9 **in** *Handbook of Semiconductor Manufacturing Technology,* Y. Nishi and R. Doering, Eds., Marcel Dekker, New York, 2000, pp. 201-86.
34. *Advances in Rapid Thermal and Integrated Processing,* F. Roozeboom Ed., Klewer Academic Publishers, Dordrecht, NL, 1996.
35. R.B. Fair, *Rapid Thermal Processing,* Academic Press, San Diego, CA, 1993.
36. M.A. Foad and D. Jennings, "Formation of Ultra-Shallow Junctions by Ion Implantation and RTA," *Solid State Technology,* December 1998, p. 43.
37. A. Atanos *et al.*,"Susceptor-Based Rapid Thermal Processing for Ultra-Shallow Junctions," *Solid State Technology,* April 2000, p. 50.
38. M. Ameen and J. Hebb, "Implanter, RTP Systems Issues for Ultrashallow Junction Formation," *Solid State Technology,* September 2001, p. 77.
39. B. Yu *et al.*,"70-nm MOSFET with Ultra-Shallow, Abrupt, and Super-Doped S/D Extension Implanted by Laser Thermal Process (LTP)," *Tech. Dig. IEDM* 1999, p. 509.
40. S. Banerjee,"High Ramp Rate Thermal Annealing for Ultra-Shallow Junctions," *Abs. Mtg. Electrochem. Soc.,* Spring 2000, Abs. 521.
41. D.M. Camm and M. LeFrancois,"Spike Thermal Processing Using Arc Lamps," *Abs. Mtg. Electrochem. Soc.,* Spring 2000, Abs. 567.
42. S. Thompson *et al.*, "MOS Scaling: Transistor Challenges for the 21st Century," *Intel Technology Journal,* 3rd Quarter 1998.
43. C. Osburn *et al.*, "Design and Integration Considerations of End-of-the-Roadmap Ultrashallow Junctions," *J. Vac. Sci. Technol.***B-18**, p. 338, 2000.
44. T. Ghani *et al., Dig. Tech. Papers, Symp. VLSI Technol.* 2000,
45. M.I. Current *et al.*, "Process Integration Issues for Doping of Ultra-Shallow Doping Junctions," *J. Vac. Sci. Technol.* **B-18**, p. 468, 2000.
46. U. Jeong *et al.*, "Devices Dictate Control of Implant-Beam Incident Angle," *Solid State Technology*, October 2001, p. 79.
47. R.N. Kleiman *et al., J. Vac. Sci. Technol.*, **B-18**, p. 2034 (2000).
48. H. Fukutome *et al.*, "Anomalous Diffusion in the Extension Region of Nanoscale MOSFETs," *Tech. Dig. IEDM,* 2001, p. 67.
49. T. Hori, "A 0.1 μm CMOS Technology with Tilt Implanted Punchthrough Stopper (TIPS)," *Tech. Dig. IEDM* 1994, p. 75.
50. M. Rodder *et al.*, "A Scaled 1.8 V, 0.18 μm Gate Length CMOS Technology: Device Design and Reliability Concerns," *Tech. Dig. IEDM* 1995, p. 415.
51. K. Miyashita *et al.*, "Optimized Halo Structure for 80-nm Physical Gate CMOS Technology with Indium and Antimony Highly Angled Ion Implantation," *Tech. Dig. IEDM* 1999, p. 645.
52. Y. Momiyama *et al.*, "Indium Tilted Channel Implantation Technology for 60 nm NMOSFET," *Dig. Tech. Papers, Symp. VLSI Technol.* 1999, p. 67.

53. Y. Taur, "25 nm CMOS Design Considerations," *Tech. Dig. IEDM,* 1998, p. 789.
54. T. Noda, "Modeling of End-of-Range (EOR) Defects for Channel Engineering," *Tech. Dig. IEDM*, 2001, p. 839.
55. T. Noda *et al.*,"Effects of End-of-Range Dislocation Loops on Transient-Enhanced Diffusion of Indium Implanted in Silicon," *J. Appl Phys.*, **88**, 4980 2000.
56. H.C.-H. Wang *et al.*,"Antimony Assisted Arsenic S/D Extension Engineering for Sub-0.1 μm NMOSFETs: A Novel Approach to Steep and Retrograde Indium Pocket Implants," *Tech. Dig. IEDM* 2001, p. 63.
57. S. Narasimba *et al.*, "High Performance Sub-40 nm CMOS Devices on SOI for the 70 nm Technology Node," *Tech. Dig. IEDM* 2001, p. 625.
58. R. Chau *et al.*, "A 50-nm Depleted Substrate CMOS Transistor (DST)," *Tech. Dig. IEDM,* 2001, p. 621.
59. M.I. Current *et al.*, "Ultrashallow Junctions or Ultrathin SOI?" *Solid State Technology*, September 2000, p. 66.

Chapter 6

ADVANCED LITHOGRAPHY I - PHOTORESISTS

Lithography is the most heavily used process technology for fabricating ICs. In 2002 some ULSI process flows use more than 20 mask levels. Thus, the entire lithography procedure is applied more than 20 times when fabricating many modern ICs. The semiconductor industry is also continuing to drive toward smaller feature sizes, and it relies on progress in lithography to help maintain this effort. Here many of the changes taking place in lithography processes in the 2001-2002 timeframe are discussed. While the basics of the lithography process are not presented in this text, such information can be found Vol. 1 (2nd Ed.) of *Silicon Processing for the VLSI Era,*[1] and elsewhere.[2,3]

Figure 6-1 shows the lithography-scaling roadmap as it was envisioned in 1998.[4] Not all of the predictions shown in this figure have come to pass (e.g., proximity-X-ray-lithography, predicted to be inserted into manufacturing by 2000, never got commercialized). Nevertheless, the data in the figure clearly indicates that many changes have occurred in lithography, and many more are occurring today. The purpose of this chapter (and Chap. 7) is to cover the changes depicted on the right-hand side of Fig. 6-1, including some that are not shown in this chart (e.g., 157-nm lithography, based on F_2-laser light-sources).

6.1 CHEMICALLY-AMPLIFIED DEEP-UV RESISTS FOR OPTICAL LITHOGRAPHY

As device dimensions were scaled to less than 1.0-μm, advances were occurring in both the DNQ/novolac resists and the lithography hardware. To achieve sub-wavelength resolution, advances in lithography were made (i.e., improvement in stepper optics and inspection techniques). Such progress allowed DNQ/novolac resists to continue to be used to fabricate device patterns in the sub-micron realm. Specifically, by using *i*-line illumination to replace *g*-line, by using *i*-line lenses with larger numerical aperture (NA) in the steppers, and by improving the novolac resin and DNQ materials, the resolution of *i*-line lithography kept being improved. Eventually, it became possible to print the features used in 0.35-μm CMOS technology with DNQ/novolac resists. Nevertheless, it was understood that new photoresists would be needed for the generations beyond 0.35-μm. Thus, a major industrial effort was launched in the early 1980's to develop photoresist systems for shorter wavelengths. This effort spanned about 10 years.

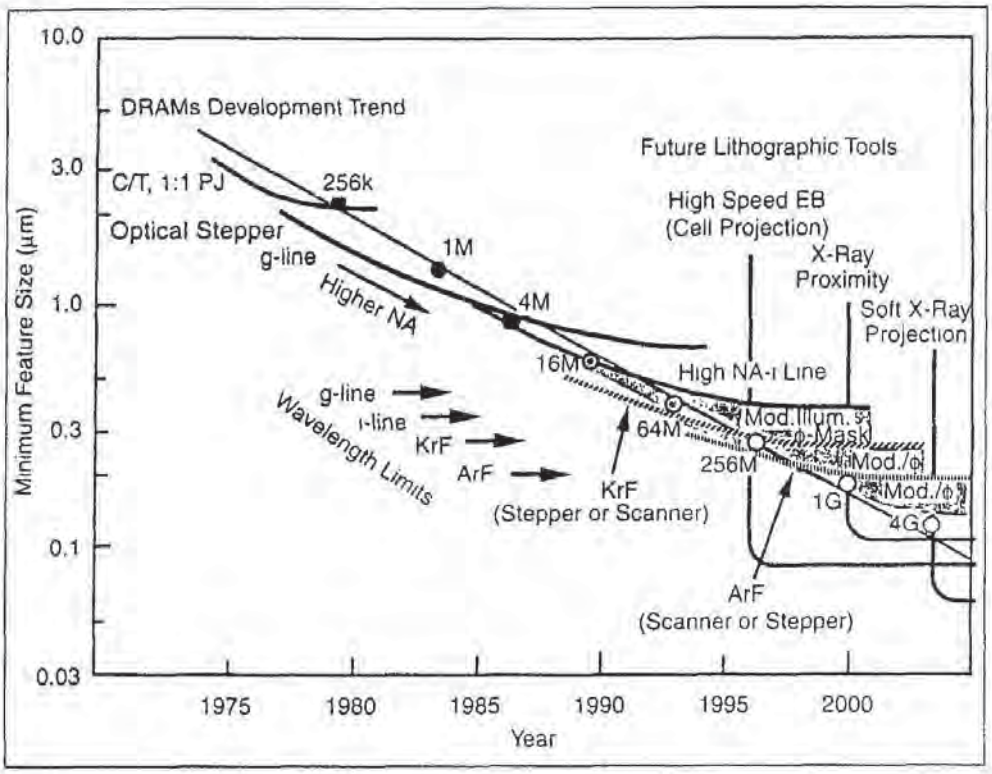


Fig. 6-1 Lithography Scaling Roadmap

6.1.1 248-nm DUV Photoresists

Eventually, by the early 1990s, when the 0.25-μm generation was reached (and used to make such ICs as 256-Mbit-DRAMs), it became clear that *i-line* lithography could no longer print 0.25 μm features (even when using such resolution-enhancement-techniques as phase-shift-masks and off-axis illumination). A shift from *i*–line to deep-UV illumination had to be made. Since both novolac resin and DNQ strongly absorb UV light at around 250 nm, DNQ/novolac resists cannot operate properly in the DUV region. Thus, the new resist chemistry developed during the 1980s (based on *chemical amplification* - CA) was adopted. This CA-chemistry also offered increased photospeed and contrast. Since CA-resists were compatible with DUV illumination, they became the resist materials of choice for advancing optical lithography toward higher resolution.[5]

In chemical amplification the important chemistry is not driven directly by the photons absorbed during the exposure step. Instead, the absorption of photon energy causes the decomposition of a *photoacid generator* (*PAG*). This causes a small amount of acid to be formed throughout the resist. (In early CA-resists such PAGs were *onium salts*.) This acid induces a cascade of chemical trans-

formations in the resist film, typically during the *post-exposure bake* (*PEB*) step. These transformations increase the solubility of the resist in the developer. Since the latter reactions are catalytic, the acid is regenerated after each chemical reaction. Thus, the same acid molecule can participate in further reactions. From 500-1000 chemical reactions can result from a single photon-absorption event. In summary, imaging in CA resists occurs in two steps: **1**) acid generation during exposure; and **2**) acid-catalyzed reactions during the PEB step.

This technology was pioneered by IBM in the early 1980s in an attempt to accommodate the low intensity of illumination produced by Hg-arc lamps at 248-nm wavelengths. Very fast resists were needed, and this led to the development of the CA reactions. Both positive- and negative-tone DUV-resists were developed. Details concerning the *chemical compositions* of these resists are described in Ref. 5. Information about the *chemical reactions* occurring during the various resist processing steps of 248-nm DUV-resists (i.e., exposure, PEB, and develop) can be found in Ref. 6.

One of the first commercially available CA-DUV-resists was a negative-tone resist (Shipley MEGAPOSIT SNR200). Its introduction allowed lithographic practitioners to gain experience with CA-resists and to witness first hand their high resolution capability. However, the SNR200-resist did not find widespread use because it is difficult to print contact holes with negative resists. Production-worthy positive-tone CA-DUV-resists soon followed, with the Shipley APEX-E resist being developed by the IBM/Shipley DUV-resist alliance. The IBM-developed CA-DUV resists have phenolic sites on poly hydroxystyrene (pHOST) which have been protected with t-butoxy-carbonate (t-BOC) groups. The IBM t-BOC resist works in positive-tone with aqueous developer, and alternately in negative-tone using an organic developer. Other major resist vendors now also offer CA-DUV-resists, including Clariant, Olin, JSR, and TOK. Positive-tone materials have become the preferred DUV resists. Some of the newer resists are compatible with TMAH developer and have been made more resistant to airborne contamination. Line/space patterns down to 165-nm, with a depth-of-focus (DOF) of 1.2-μm can be printed with such resists. Negative-tone CA-DUV-resists are also available, although not as many types are offered. Because of the improved aerial image of dark-field masks versus bright-field masks, the ability to print isolated-lines with negative resists is superior to that of positive resists. Thus, for some pattern types, negative resists are preferable. CA-DUV resists, however, are expensive. Their cost in 1998 was \$1500–\$2000 per gallon, which is about 2–3 times higher than DNQ/novolac resists. Future CA-DUV resist formulations are predicted to be even more costly.

6.1.1.1 Amine-Contamination of DUV CA Resists: While the biggest advantage of CA-resists is their high sensitivity (i.e., fast photospeed, in the range of 10–50 mJ/cm^2), the catalytic nature of the CA process makes reproducibility a concern. If the catalytic chain of reactions is somehow interrupted, many of the lithographically important reactions will not occur, leading to catastrophic resist failure. That is, prior to post-exposure-bake (during which the photoacids perform the task of deprotection), these acids are susceptible to neutralization by bases.

The most common mechanism that quenches the catalytic reactions occurs as a result of *ambient contamination.*[7] Specifically, the culprits are trace quantities (on the order of 10 ppb) of vapors of volatile base compounds (such as amines, ammonia, ammonium hydroxide, HMDS, or N-methylpyrrolidone [NMP - a common ingredient in strippers]). Such trace levels of amines in the ambient air can significantly affect the photochemistry because (as noted above) only a small quantity of photo-acid is produced by the exposure dose. If the ambient contains any of these contaminant vapors, the top surface of the resist will absorb them during the time that elapses between the exposure and post-exposure-bake (PEB) steps. These base molecules will then react with and neutralize the photo-generated acid. The surface region of the resist is now devoid of acid, and thus becomes less soluble than the acid-containing regions below. This less-soluble skin develops more slowly, resulting in the formation of a T-top structure (Fig. 6-2).

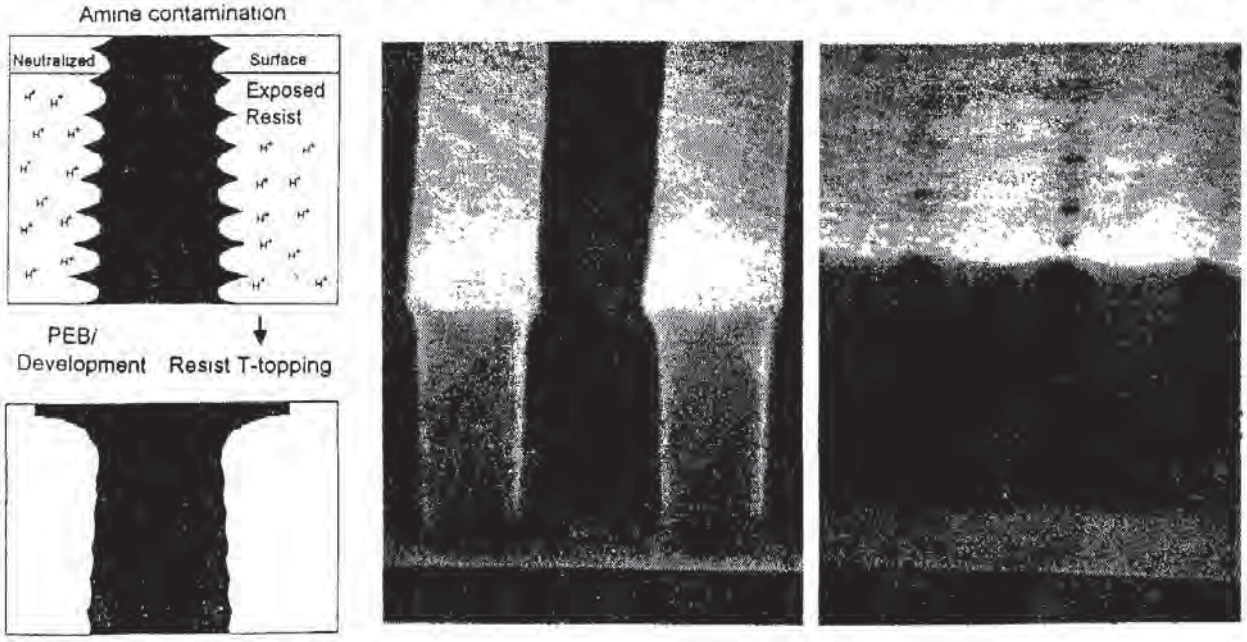


Fig. 6-2 (a) Environmental amine contamination of a chemically amplified resist, resulting in "T-top" formation after development.

This causes linewidth variations from wafer-to-wafer, as the extent of the skin formation is not controlled. In addition to airborne contamination, the acids also decay with time. Thus, any delay between exposure and PEB exacerbates the linewidth control problem. After PEB, resist susceptibility to amine contamination drops to negligible levels.

Let us elaborate on the problem of T-top formation. As noted above, the resist is most sensitive after the coated wafer is exposed in the exposure tool and transported back to the PEB/development tool, until the time it is baked (PEB). Typically the time delay between flash and post-exposure bake may be less than 3 minutes, depending on tool loading, but on occasion it may be 10 minutes or longer. A 10-minute pause can easily cause a linewidth variation on a wafer that may have just finished exposure but is not yet baked. If this event happens when a wafer is in an area where the amine contamination level has reached 5 ppb (or more), it is possible to cause a CD shift of as much as 10-20-nm. If the CD shift is great enough, changes in the resulting linewidth can cause the electrical characteristics and performance of the devices to change (possibly to the point where the device will not perform to specification and its speed or function is diminished). The length of time the wafers are exposed relative to the contamination concentration determines the potential risk to linewidth control. The effect of amine contamination on an example DUV resist is shown in Fig. 6-3 for three conditions: **1**) no time delay between exposure and PEB; **2**) 45-minute delay between exposure and PEB (with the wafer being held within the scanner enclosure, in which the amine concentration was about 1 ppb); and **3**) 45-minute

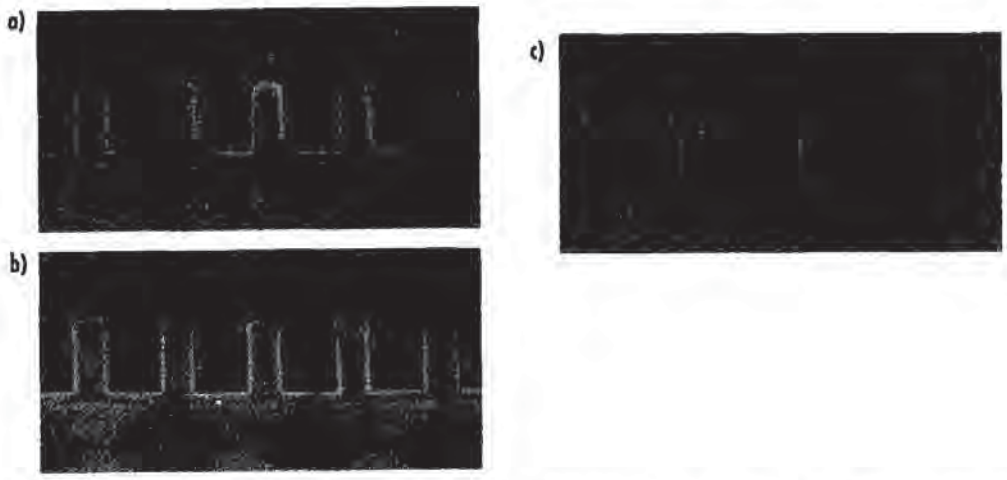


Fig. 6-3 The effect of molecular bases on a DUV-resist is shown with: a) No delay between exposure and PEB; b) after a 45-min delay in the exposure tool; and c) after a 45-min delay in a cleanroom ambient.

delay between exposure and PEB (with the wafer being held in the cleanroom-ambient which has an ambient-amine-concentration of about 5-ppb). Significant T-topping occurs as a result of allowing the wafer to remain in the cleanroom-ambient for 45 minutes.

As shown above, the environmental contamination problem can be quite serious, as trace quantities of basic (i.e., as opposed to acidic) vapors are ubiquitous in the semiconductor fabrication ambient (typically in the 5-10-ppb range). Several methods have thus been established to combat this problem. The first involves filtration of the cleanroom air with activated charcoal. This was the practice first adopted by the industry when DUV-resists were initially being used, but it only reduces the nominal amine concentration in the cleanroom ambient to the 5-10-ppb range. (Environmental base contamination can also be neutralized and further reduced by adding weak acids to the filters.) More recently (as noted above), scanners and track systems (where the resist-coating, exposure, and development steps are performed) are housed in enclosures, and the air that enters these enclosures is purified with charcoal filtering and monitoring systems. Such airborne-amine-filtration and real-time monitoring systems are built by Extraction Systems (Franklin, MA) and other firms (see Figs. 6-4 and 6-5). According to the ITRS Roadmap, the total airborne contamination levels inside such DUV-lithography-enclosures must be kept below 1-ppb (molar) by such filtration systems as long as CA-resists are used. Another approach is to have the wafer track place the wafers in "safe holding" locations after exposure. Other ways to controlling this problem include: **a**) coating resist layers with a protective water-soluble transparent polymeric film (and removing this "sealing" film prior to development with a water rinse); **b**) incorporating stabilizing additives to the resist; and **c**) reducing the rate of absorption of contaminants by "densifying" the resists through annealing.

Nevertheless, even use of all of the above techniques is not completely effective in curing the problem of amine-poisoning of CA resists. Therefore, it is important that additional measures be employed to minimize the levels of amine in the cleanroom ambient. This is a challenging task since there are many potential sources of amines in a fab. Amine levels in wafer fab areas can easily exceed the nominal 10-ppb concentrations when small quantities of developer or adhesion promoters (HMDS) are spilled. (Note that HMDS application is an ammonia- producing process that therefore requires ample exhaustion to prevent

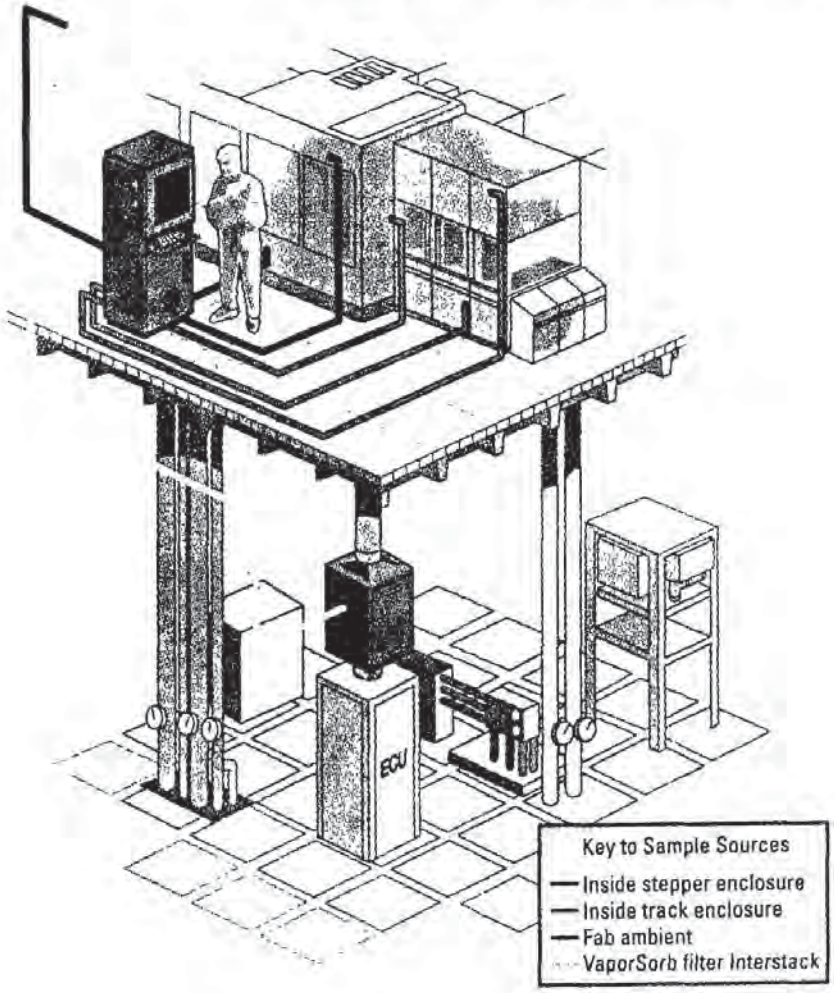


Fig. 6-4 Schematic of filtration units connected to photo-island clusters to avoid amine contamination of chemically amplified resists. Courtesy of Extraction Systems, Inc.

leakage into the track environment, causing amine levels within the track system enclosure to rise above 1-ppb.) Many paints, plastics, and epoxies also outgas amines. Materials that initially evolve large quantities of volatile bases, but quickly reach low-outgassing rates can be used in the fabrication of lithography tools. However, their use can be problematic in operating a wafer fab. Some materials continue to evolve amines over long periods of time and should not be used in wafer fabs, nor in the construction of lithography equipment.

6.1.1.2 Current State of 248-nm Resist Technology: Devices fabricated at 0.35-μm use a mixture of optical lithographies, including traditional *i*–line and 248-nm-DUV (the latter employing single-layer CA-resists). The conversion from *i*–line to 248-DUV was completed in the transition to 0.25-μm devices. The

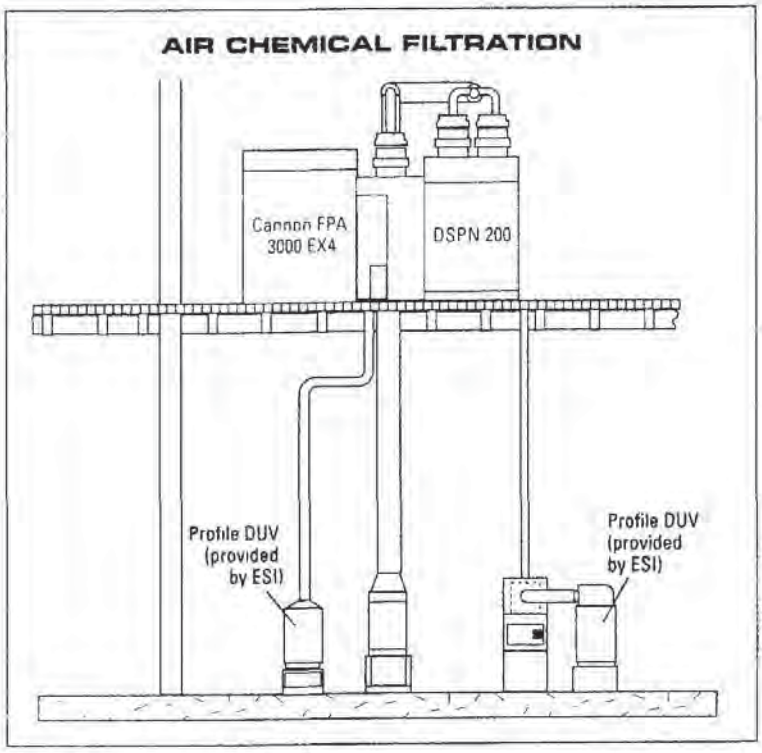


Fig. 6-5 Cabinet-style chemical air filtration was added to the existing integrated filters to provide protection from possible spills and elevated room concentrations.

resolution limits for 248-nm resists continues to decrease. By adding such enhancements as phase-shifting-mask technology and/or more complex multi-layer resist schemes, the use of 248-nm-lithography is predicted to be extended to features as small as 0.13-μm. In fact, some claim that positive resists for 248-nm can image down to 0.1-μm (100-nm) for isolated lines in logic applications, and 130-nm in dense arrays using the latest high-NA (0.8) scanners. (Smaller isolated lines can be imaged than can features in dense arrays because scattering-bar and OPC technologies are more readily applied to isolated structures.) However, at these critical dimension limits, it is very hard to maintain adequate process windows in a production environment and process optimization is engineering-intensive. The applications in which 248-nm will be pushed to 100-nm will therefore be very limited.

With manufacturers supplying fifth and sixth generations of 248-nm resist, these resist types are maturing rapidly. For example, their etch resistance is now comparable to *i*-line resists. The 248-nm resists are also optimized for specific challenges, resulting in resists for isolation layers, contact holes, metal lines, etc. Although such layer-specific resist specialization causes complexity in the fab line, it is mandatory in order to achieve sub-wavelength lithography.

TABLE 6-1 Linewidth Sensitivities of Commercially Available DUV KrF Resists

Resist	Supplier	Linewidth sensitivity (nm/°C)
Apex E	Shipley	16.0
UVIIHS	Shipley	7.5
UV6	Shipley	2.6
TM-461	JSR	2.6
DP-024	TOK	1.8
ARCH 2	Arch	~0

Another issue involving 248-nm DUV-resists is their sensitivity to post-exposure-bake temperature. The earliest commercially available 248-nm resists (e.g., APEX, from Shipley) exhibited large linewidth variations with fluctuations in the temperature of the PEB (16.00 nm/°C). However, as shown in Table 6-1, resist suppliers responded to industry requests to provide better linewidth control. (For example, Shipley later produced the UVII/HS resist - and thereafter the UV6 resist - which progressively improved this parameter.) However, these linewidth sensitivities have driven hot-plate specifications, since the resists with essentially no post-exposure linewidth-variation may not be universally applicable to all feature types, nor are they necessarily able to provide the best resolution.

In order to satisfy the extraordinarily tight linewidth-control requirements of today's and future processes, hot-plates must therefore exhibit excellent across-plate uniformity – significantly less than 0.5°C (range) – and must be matched to tighter levels. This is a challenge not only in terms of achieving the required capability, but it also poses a formidable metrology challenge. Resistance temperature detectors are the best available thermometers today. These are usually bonded to a silicon wafer, which is then placed on the hot plate. This placement is also difficult, since the wires must go between the wafer and the measurement electronics.

As noted above, printing isolated lines is easier than printing nested features. Consequently, contact-hole layers have rapidly become the ones that are most difficult to image (and may be the first to transfer to 193-nm). One example of how layer-specific resist specialization is being used to tackle the contact layer imaging problem is the so-called *RELACS* process.[8] That is, this is a process developed by Mitsubishi Electronic Corp. for shrinking KrF-imaged via holes

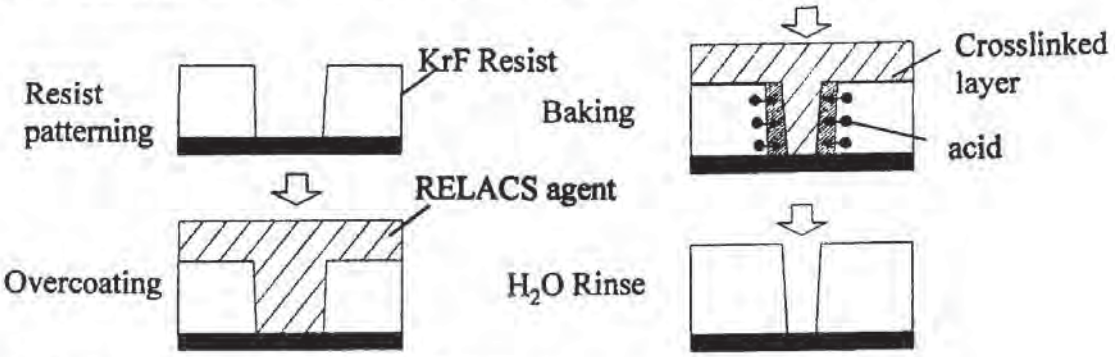


Fig. 6-6 Process flow of RELACS.[8] (© IEEE 1998)

from 0.2-μm to 0.1-μm using a coat, diffusion-bake-and-rinse sequence, following resist-exposure and develop. That is, after the coating step, acid from the unexposed resist diffuses into the wall of the via hole covered by the coating – and then into the coating material – during the bake step. The coating material is crosslinked by this acid (see Figs. 6-6 and 6-7). After a rinse step, some of the coating material remains behind, thereby causing the opening size of the via to shrink. Clariant AZ has licensed the materials from Mitsubishi (R200 coating and R2 developer) and commercializes the process as RELACS (Resist Enhancement Lithography Assisted by Chemical Shrink).

Once the performance of single-layer 248-nm DUV-photoresists based on pHOST chemistries becomes exhausted, bilayer-resists may be adopted to extend the 248-nm-based lithography process. Bilayer processes decouple requirements for imaging, performed by the thin (200-nm thick, or less) top resist, from needs

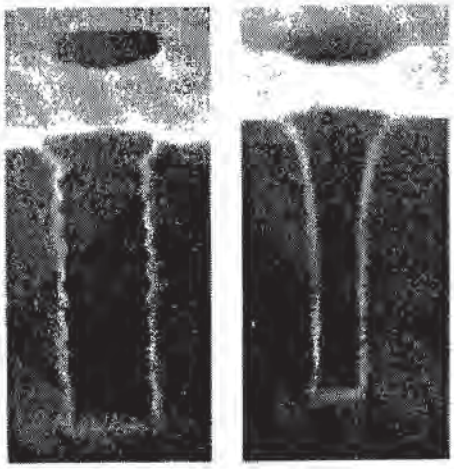


Fig. 6-7 The RELACS process shrinks these KrF-imaged 0.22 μm contact holes (left) to 0.11 μm (right) using a coat, diffusion bake, and rinse step after wafer patterning. (Courtesy of Clariant AZ)

for etch resistance, managed by the thick (400-1000-nm thick) bottom resist. The top layer uses standard single-layer resist imaging, and the image is transferred down anisotropically using O_2/SO_2 plasma etching. The bottom layer acts as an anti-reflecting coating (ARC), reducing the pattern-width degradation from underlying topography. Importantly, bilayer-image performance tends to be more limited by the capability of the scanner. At least one bilayer approach has been implemented in manufacturing. The CARL (chemically amplified resist lines) developed by Infineon Technologies, is being evaluated at 248-nm and 193-nm. (See Vol. 1, 2nd Ed. for more details on the CARL process.)

Another bilayer approach is being proposed for dual-damascene patterning with 248-nm lithography. Arch Chemical offers a thin imaging system (TIS-2000®) optimized for dual-damascene patterning. The top resist layer provides patterning and the bottom organic-layer acts as the BARC, while also planarizing via layers for trench imaging in a via-first approach (Fig. 6-8, see also Sect. 15.9.1.5).[9] One advantage to this process is its dual use at 248-nm and 193-nm.

6.1.2 193-nm DUV Photoresists

The insertion point for 193-nm lithography is predicted to be the second-generation 0.13-μm or even 0.10-μm devices (late in 2002 or early 2003). Successful extension of 248-nm resist is stalling the 193-nm transition. In addition, the compressed timeline for arriving at the 0.13-μm node on the SIA Roadmaps is also part of the reason. That is, if there were four years allowed for developing the 0.13-μm technology, then 193-nm lithography would have a good chance of being accepted into manufacturing from the beginning of that technol-



Fig. 6-8 In the via-first damascene approach, the underlayer resist serves the purposes of planarization, BARC and uniform via fill to protect the via bottom during the trench etch. (Courtesy Arch Chemicals and IMEC)

ogy node. But, since there is not that amount of time, there is more of an emphasis being placed on extending 248-nm lithography.

At 0.13-μm there is substantial benefit to transferring to 193-nm using standard binary reticles, rather than using 248-nm and enhancing the reticles with RET. However, the decision to make this transition will depend on the maturity of 193-nm resist materials and the performance of 193-nm exposure tools with very high NA (i.e. > 0.68).[10] In addition, at the 0.13-μm node, economics (rather than technology) may influence the insertion point of 193-nm lithography. That is, some customers (who manufacture a large number of parts from a single mask set) may try to extend the use of 248-nm lithography by any means possible, including the route of using more expensive reticles. Others (such as foundries, which on the average, run fewer wafers per mask) may prefer the 'simpler" approach of driving to a shorter (193-nm) wavelength. It is estimated that possibly 10 mask levels will require the new systems (enhanced-248-nm or 193-nm). Critical masks will include gate, active area, contact, first metal and first via..

One of the most significant aspects of 193-nm resist development is that after nearly a decade of R&D (primarily performed empirically, due to the complexity of reactions within the resist), a single superior polymer backbone material has not yet been identified (even though the insertion point of 193-nm lithography is drawing closer). That is, no single polymer platform material seems to be able to provide the best process for every application. Instead, there are three seemingly-viable resist polymer types, each exhibiting some benefits – but none of which have all the benefits needed to be used across all production lines. What is predicted to occur is that users will select different chemistry platforms for different applications.

That is, despite the fact that process engineers prefer one resist/one supplier, at 193-nm there are three useful polymer platforms: **1**) *acrylates* (acrylate platforms with attached ali-cyclic structures); **2**) *pure cyclic olefins*: and **3**) *cyclic olefin/maleic anhydrides (COMAs)*.[10] (These are all also different than the hydroxy-styrene-based resists used for 248-nm lithography.) The acrylates for example, reportedly provide the best dense-line performance (particularly with the adamantane methylacrylate chemistry developed by Fujitsu), but they generally exhibit poor etch-resistance. On the other hand, the cyclic-olefin polymers offer better etch—resistance. IBM (in collaboration with Arch Chemicals) is pursuing this group of materials (which reportedly show intrinsically higher robustness toward fairly harsh etch steps). Finally, Clariant AZ licensed a free-radical copolymer of func-tionalized cyclo-olefins and maleic anhydride from Hyundai. This polymer appears to be very cost-effective and may

migrate back into 248-nm to reduce resist cost.

The partial benefits that each polymer offers has caused resist suppliers to pursue multiple platforms for 193-nm resists. That is, by 2000 many resist suppliers were working on two or three of these different polymer families (whereas in 1997 each 193-nm resist-supplier had its own, single platform). In addition, a lithography consortium headed by IMEC has been established to evaluate various 193-nm resists from Arch Chemical, Clariant, JSR, and Shipley on a common lithography tool (the ASML 5500/900 scanner with integrated TEL track). This will simplify comparisons of resists whose performance is often measured under different conditions. By 1999 several promising 193-nm resists were emerging (although at that time they still remained 1-2 years from being production worthy). Nevertheless, using a single-layer 193-nm resist, a 0.6-NA scanner and alternating phase-shift mask, 70-nm gates could be patterned (Fig. 6-9).[21]

Beyond selecting the most appropriate polymer platform, resist chemists are charged with finding the photoacid generator (PAG) best able to deliver strong dissolution inhibition, high photospeeds, thermal stability, and tailorable photoacid size (the last factor correlates strongly to imaging performance).

6.1.2.1 Problems of 193-nm Resists: As noted above, the 193-nm photoresist will need to be selected by chip manufacturers in 2001-2002 to begin developing 193-nm lithographic processes. Together with the issues of selecting resists based on their polymer platform, a set of other resist performance characteristics also needs to be considered. We discuss those issues here.

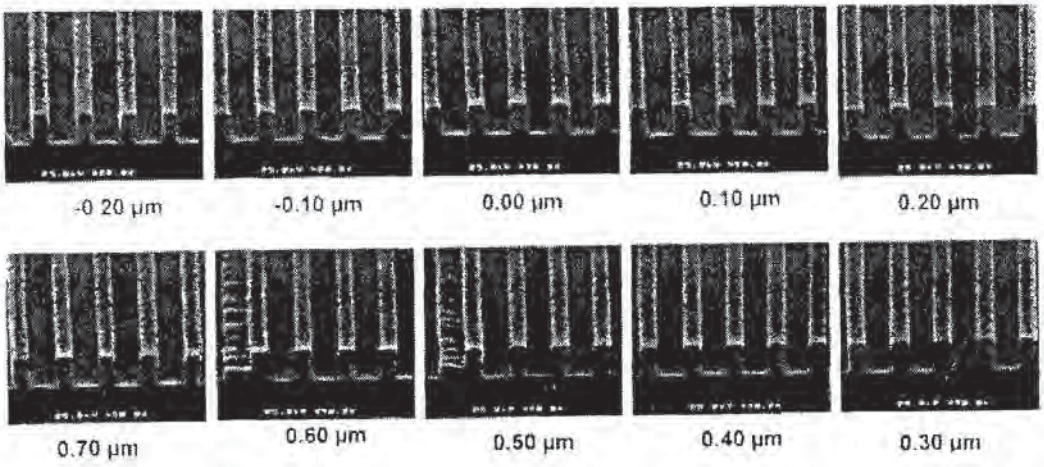


Fig. 6-9 70-nm gates patterned with 193-nm resist and alternating phase-shift mask. Courtesy of Semiconductor International.[21]

The *primary performance factors* involving the still-immature 193-nm resists are those involving process latitude, namely: exposure-latitude and DOF (i.e., an exposure latitude of ~20% and a DOF of 0.5-μm are needed at 193-nm). The are also numerous *secondary performance characteristics* including: etch-resistance; degree of linewidth variation with fluctuation in post-exposure-bake temperature; line-edge-roughness; resist-collapse; resist adhesion (lifting); resist shrinkage during irradiation with e-beams during CD inspection; and resist outgassing.

Etch resistance becomes more critical for 193-nm resists because thinner resist layers will have be used than in previous lithography generations. Resist thickness is expected to be reduced to 400-600 nm in 193-nm (down from about 1000-nm in prior generations of lithography). The advances in etcher design and the use of high-density-plasma systems are helping to improve 193-nm resist thinning during etch steps. However, in cases of insufficient dry-etch-resistance, use of a hard-mask instead of resist-masking may need to be considered.

The *degree of linewidth-variation* in the face of post-exposure-bake temperature-fluctuation is another inherent problem of the immature 193-nm resists. By 2001 193-nm resists exhibited linewidth variations of 10 nm/°C, which is significantly higher than that exhibited by the latest generation of 248-nm resist at that time (see Sect. 6.1.1). But, the same situation existed when the early 248-nm resists were first released. The target value for 193-nm is 3 nm/°C.

Line-edge-roughness is much more pronounced at 193-nm than at 248-nm. Complex interactions occur between lithographic performance, line-edge roughness, and etch-resistance. Resist chemists are seeking the right compromise. Some researchers claim that line-edge-roughness (LER) can be reduced to the 5-7-nm range. The effect of LER on device performance is still being quantified.

Resist-collapse becomes a major issue as the resist features are brought closer together. That is, as developer and rinse water are being removed from developed wafers, surface-tension forces cause closely-spaced adjacent lines to be pulled together (Fig. 6-10). This phenomenon limits the aspect-ratio of resist features (another reason why resist films have smaller thicknesses as features are scaled). Resist-collapse becomes common when resist-feature aspect-ratios exceed about 4:1. There is some concern that resist-collapse may be one factor that could delay the deployment of 193-nm lithography.

The immature 193-nm resists are also prone to *lifting* since their formulations are not yet optimized for adhesion (as are those of *i*-line and 248-nm resists). Adhesion can be optimized with respect to wafer-temperature and HMDS-delivery for each resist type. Lifting can also be addressed by reducing the resist thickness. Using a multilayer resist is another potential remedy for resist-lifting.

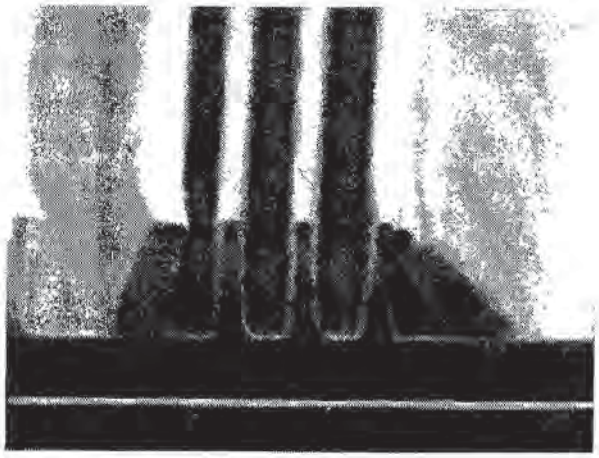


Fig. 6-10 Closely spaced, high-aspect-ratio lines of resist that have collapsed after development.

The 193-nm resists also exhibit significant *shrinkage* (both in the vertical and lateral directions) when they are irradiated by electron breams (as occurs when CD inspections are performed using an SEM). Such resist shrinkage was also observed in early 248-nm resist, but modified resist formulations helped to overcome that problem.[11] Some 193-nm resists exhibit more shrinkage than others, but the tradeoff is that those that have better shrinkage behavior tend to have worse photospeeds and are more prone to resist collapse. Additional work is needed to make 193-nm resists able to meet all of these needs.

Another significant problem with 193-nm resists is that they *outgas organic compounds* that can then build up on scanner optics. This issue may be addressed by PAG selection or by using various additives. Aromatic fragments such as the iodobenzene liberated by iodonium salt sensitizers are among the most critical outgassing species. While the outgassing problem is evident with 248-nm systems as well, the working distance between the lenses and resist at 193-nm is shorter, and the adsorption of vapors is greater at shorter wavelengths. Thus, it is predicted that the outgassing problem will further increase for 157-nm systems.

6.1.3 157-nm DUV Photoresists

The existing deep-UV lithographic technologies and associated optical extensions combined with early 193-nm-based systems are expected to provide sufficient resolution to allow them to be used for the 130-nm device generation outlined in the 1999 SIA Roadmap. However, extension of optical lithography

beyond that point may require that 157-nm DUV-illumination systems be adopted by 2005. If this time-schedule is to be met, the first pilot-line quantities of 157-nm resist will need to be delivered to SEMATECH by the first-half of 2002 to begin process development efforts toward the 100-nm device node. Volume production capability using 157-nm illumination systems should be then become available by the 1st quarter of 2005.[12]

The question of "Will 157-nm lithography be ready by 2005?" was debated at the SPIE Microlithography Conference in March 2001. However, there was no consensus as to whether the schedule will (or even, should) be met. Nevertheless, considering that the development cycle for 248-nm DUV resist materials spanned at least 10 years (1980-1990), and the 193-nm resist development program was approximately 8 years long (1991-1999), the 157-nm resist development efforts are faced with a much more abbreviated and challenging development schedule. In addition, the cost of developing a 157-nm resist is also expected to be very high (estimated to be $50-60M). Since such a financial burden is not feasible for any single company, the International SEMATECH formed a University-based consortium to perform fundamental research on 157-nm resists. Just as significantly, IMEC and ASML have begun a cooperative 157-nm lithography development effort, with funding support from many top device manufacturers, including AMD, Infineon, Intel, Micron, Philips, Samsung, ST Microelectronics, and UMC. The First International Symposium on 157-nm Lithography (sponsored by 43 companies), was held in Dana Point, CA in May 2000, and the Second Symposium was held in the same location in May 2001. In any case, researchers are confident that the industry will design workable resists for 157-nm lithography, but they are less sure that they will be able to do it by the 2003 deadline (i.e., set by the possibility that this will be time at which it will be necessary to pattern critical levels in the 70-nm device generation).

The biggest difficulty facing 157-nm lithography is that the illuminating radiation consists of such high-energy photons (7.9 eV). Most polymers used to make up the backbone of previous generation resists strongly absorb such high-energy photons. Thus, it is difficult to find polymers for 157-nm resists that have sufficient transparency (as well as adequate etch resistance, and good adhesion to substrates). Fortunately, however, two promising candidates for the *backbone* (or *platform*) polymers in 157-nm resists have been identified: **1**) highly fluorinated hydrocarbons (fluoropolymers based on norbornane); and **2**) silicon-based siloxanes. A key challenge to developing 157-nm resists based on the latter group of materials (siloxanes) is the ability to achieve a sufficiently high glass-transition temperature.

Fundamentally, the design of materials for developing 157-nm resists is identical to that used to develop resists for the previous DUV lithographic generations. That is, the materials must satisfy the following set of requirements common to all CA-photoresist materials: **1**) high transparency at the exposure wavelengths; **2**) sufficient resistance to plasma-etching processes; and **3**) functional units that are capable of undergoing efficient photochemical transformations that change their solubility in developer solutions. For positive-tone resists, this is typically accomplished by designing the matrix-resin (backbone material) with an acidic functionality that is masked by an acid labile protecting group. Strong acids generated within the exposed regions of the photoresist-film catalyze the thermolysis of the protecting groups, rendering these regions soluble in the aqueous-base-developer solutions. The resists should also have glass-transition-temperatures higher than the processing temperatures, acceptable shelf-storage lifetime, and minimum toxicological risk. A single-layer-photoresist process for imaging at 157-nm is also clearly preferred over a multilayer-resist approach.

When attempting to develop the next-generation of resists, it is often convenient to adopt a modular approach for the design of new photoresist materials. This allows the polymer design to be broken down into several functional sub-units that may be addressed in a somewhat independent fashion (Fig. 6-11).

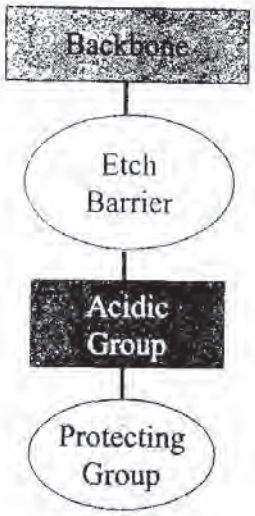


Fig. 6-11 A modular approach in the design of new photoresist materials provides flexibility in design, but transparency among all subunits must be maintained.

Since the 248-nm and 193-nm resist materials are too-absorbing at 157-nm, and since the 157-nm resists were still in their earliest stages of development in 2001, a two-pronged approach was chosen to develop the technology of 157-nm lithography. The first prong involved using existing resists, but at a greatly reduced thickness (60-70-nm-thick). This provides a pathway for tool-testing and for developing early baseline-processes. The second-prong pursues development of the 157-nm resist itself (i.e., exploring the use of novel polymers - such as the fluorinated variations of the resins used in 248-nm resists). In these research efforts, attempts are also made to optimize the formulations and imaging conditions required to print sub-100-nm features. This is a complex task, since it is necessary to explore a variable space that includes the photoacid-generator, base, dissolution-inhibitor, base-polymer, exposure, bake, and development. Figure 6-12 shows some preliminary results of 70-100-nm-features printed using early formulations of 157-nm resists that have evolved from that development effort. Imaging is performed at 0.6-NA, 50-mJ/cm^2, with an alternating phase-shift-mask, using 148-nm-thick resist on a 70-nm antireflective-layer (DUV30), followed by a 60-sec 140°C PAB, a 90-sec 130°C PEB, and development for 20-sec in a 0.26-N TMAH developer. A good image quality is obtained, with sidewall angles between 85-89°.

Beyond having to meet the performance requirements for reduced line-edge-roughness, fast-photospeed, dry-etch-resistance, and solubility in aqueous-base - developer, 157-nm resists have a strong tendency to outgas materials, which settle on the lens. At 157-nm, even a monolayer of oxygen or water can significantly change the transmission of light through the lens and even attenuate light. Fluorocarbons are less of a concern than silicon dioxide, because the user can bleed oxygen into the system with the laser on, making the fluorocarbon residues volatile. With siloxanes, however, SiO_2 deposits are more difficult to remove. Attention to designing effective gas (nitrogen) flows between the wafers and lenses in 157-nm lithography tools will help combat this problem.

6.2 ANTI-REFLECTIVE COATINGS (ARCs)

Interference effects caused by the reflection of light that is transmitted through the photoresist and reflected off the substrate and back, can interfere with incoming light and cause the resist to be unevenly exposed. (See Vol. 1, 2nd Ed., p. 517-24 for a discussion of such effects.) Instead of reducing such interference effects through the modification of the resist itself (as also discussed in Vol. 1), methods that reduce reflectivity at resist interfaces can provide linewidth control with minimal loss of resist performance. That is, to dampen these reflections,

anti-reflective coatings (*ARCs*) can be used.[13] By minimizing standing waves and reflective notching from substrate reflections, ARCs provide improved resolution and better CD control.

ARCs suppress the reflections either by: **1**) *attenuating light* that passes through the ARC or; **2**) by *matching the index of refraction* of the ARC to the resist-system at the exposure-wavelength employed (such that destructive interference of the light passing through the ARC occurs). These layers can be deposited either on top of resist layers (forming a *Top AR* coating or *TAR*), or on the surface of the wafer beneath the resist (to form a *Bottom AR Coating* or *BARC*). For 0.25-μm lithography (and below), BARCs must be able to suppress more than 99% of the substrate-reflected-light to provide sufficient CD control. Although it significantly increases the process complexity, use of *both* a BARC and a TAR have been reported for layers that require the tightest CD control (i.e., poly gate patterning).

Bottom anti-reflective coatings are materials applied between the substrate and the resist. They are used extensively in applications such as gate definition and for opening contacts, where it is important to keep tight control over critical dimensions. They are not used much at the interconnect level, since the layer of TiN typically used to control electromigration and prevent hillocking on aluminum lines has the extra benefit of being a good anti-reflective coating. As the industry goes to shorter wavelengths, the challenges of minimizing reflections increase. That is, the reflectivity of the substrate becomes higher with

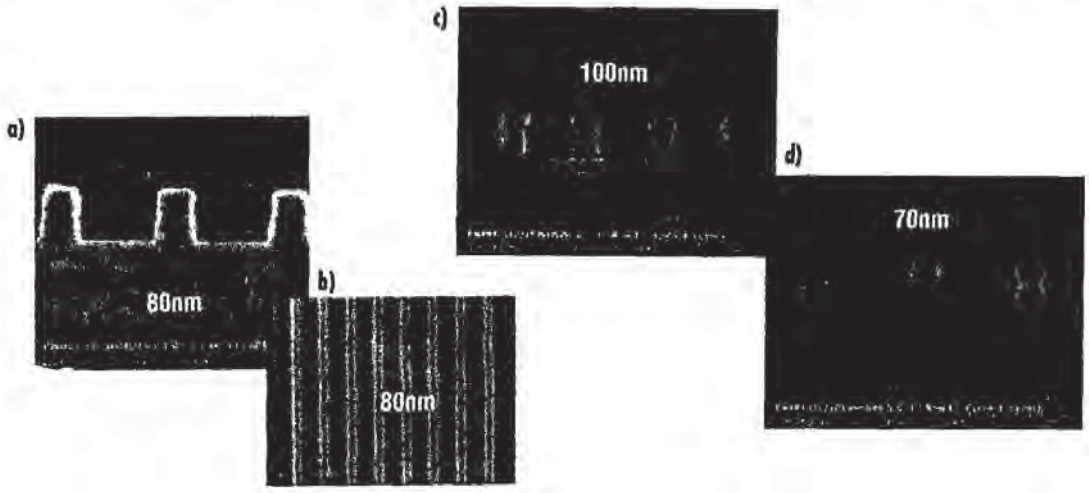


Fig. 6-12 Using 157-nm exposures, phase-shift masks, and 148-nm thick resist on a 70-nm AR layer, various resist recipes resolve at linewidths of: a) and b) 80 nm; c) 100 nm; and d) 70 nm.

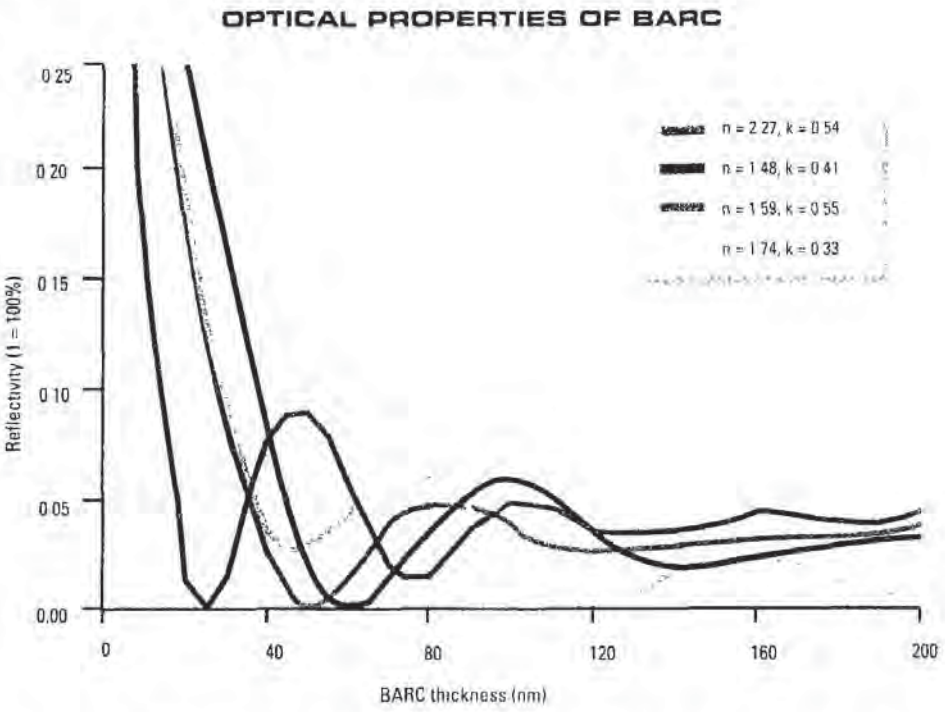


Fig. 6-13 Reflectivity is a function of film thickness, refractive index (n) and extinction coefficient (κ).

decreasing wavelength, and thus interference problems also become more severe. At the same time, advances such as phase-shift-masks, off-axis-illumination, and optical-proximity-correction have allowed features to be printed that are smaller than the wavelength of light. This has made the elimination of CD variations caused by reflected light to be even more critical.

BARCs reduce the problems involving reflectivity by decreasing the value of the reflection coefficient of the resist-substrate interface. They can accomplish this in two ways. First, the BARC may absorb the incident light that manages to penetrate to the bottom of the resist (*absorption*). This requires that a material with a high extinction-coefficient κ be chosen for the BARC, and/or that a thick BARC layer be used. Unfortunately, if the value of κ is too high, the reflectivity of the resist/BARC interface increases – which degrades the effectiveness of the BARC. Thus, useful values of κ are typically in the range of $0.22 \leq \kappa \leq 0.5$ in order to maintain reflectance back into the resist below 10%. For processes that used *i*-line resist, a fairly thick BARC-layer was used (60-90-nm thick, see Fig. 6-13).[13] But DUV-lithography mandates the use of thinner resist and BARC layers because of decreased depth of focus. One of the major challenges for BARCs is maintaining a small enough reflectivity as the BARCs get thinner with decreasing feature size (see also Fig. 6-13).

Thus, a second approach which decreases the value of the reflection coefficient of the resist-substrate interface is favored for 0.25-μm lithography (and smaller). That is, a BARC material whose index of refraction and thickness match the incident light wavelength is sought (*index-matching*). To implement index-matching, a structure is formed that has an optical-path-length of $\lambda/4$ for the wavelength of light being employed. Such a $\lambda/4$-optical-path-length will exist in a thin film (the BARC) made of a material having the correct index-of-refraction and thickness (in the range of 50-80-nm). Light strikes the BARC/resist interface, and some is reflected back toward the air/resist interface, but the remainder enters the BARC.

The light entering the index-matched BARC layer will pass through the BARC layer twice (i.e., once after reflecting from the BARC/substrate interface). These two passes through a quarter-wavelength optical-path-length will produce a light wave with a half-wavelength phase-shift directed toward the resist/BARC interface. This phase-shifted wave interferes destructively with the part of the light originally reflected from the BARC/resist interface. Thus, reflectivity at this interface is decreased. Complete destructive interference only occurs if the amplitude values of the interfering waves are identical. Tight control of BARC thickness (to ensure the two-pass optical-path-length is exactly $\lambda/2$) is an obvious requirement for index-matching by the BARC layer. If thicker BARCs can be tolerated, further reduction in reflectivity will occur by making the index-matching BARC thicker (i.e., to a two-pass optical-path-length of $3\lambda/2$).

In fact, all BARCs employ some of both the absorber and index-matching mechanisms. In any case, if the BARC is effective, little of the light incident on the resist/BARC interface is reflected back toward the resist/air interface, and as a result, the *clearing-dose swing ratio* (see Vol. 1, 2nd Ed., p. 522) is smaller. Standing-wave-effects are reduced because the smaller amplitude of the reflected light decreases the standing wave intensity. Since less light scatters off the bottom surface, the severity of notching is also mitigated. BARCs thus compensate for variations in wafer topography, substrate reflectivity, and surface characteristics. One price paid for these benefits is increased process complexity (and, of course, increased process costs, in terms of equipment and materials). For example, a separate process may be needed to remove the BARC edge-bead. Another issue which arises when BARCs are employed is that their use may also significantly extinguish the light signals that come from alignment marks (these signals arise from the scattering of light from the mark). A variety of deposition techniques have been used to deposit BARCs including spin-on, PVD, and CVD processes.

6.2.1 Organic BARCs

Spin-on organic BARCs are primarily absorber-type polymer films that contain highly absorbing dyes (introduced into the polymer at concentration levels that deliver appropriate extinction-coefficient values). They must also be non-bleaching at the exposure wavelength. Such *organic BARCs* are applied directly to the substrate to a thickness in the range of 60-120-nm. The resist is then spun on top of the BARC, using conventional track equipment. Suppliers of organic spin-on BARCs include Brewer Science, Clariant, Shipley, Hitachi, and Tokyo Ohka. Proponents of the organic approach point out that organic BARCs are advantageous in terms of cost-of-ownership, refractive-index-reproducibility, planarization-capability, film-thickness-tolerance, rework-capability, and surface-control.

Early organic BARCs were about 200-nm-thick, and were wet-etched (in a step often called "ARC open"). Since wet etching is an isotropic-removal process, the resulting undercutting of the BARC layers made them unsuitable for optical lithography smaller than 0.5-μm. Newer organic BARCs are designed to be developed by an anisotropic dry-development process using an oxygen plasma. This extends their use to DUV lithography. However, the self-planarizing nature of spin-on films does not make the dry-developing process a complete solution to the above undercutting problem. That is, when spun-on over topography, some regions of the BARC will be thicker than others, and this causes a problem when organic BARCs are dry-developed. An anisotropic dry-etch-step can be used to etch the BARC in the flat areas of the wafer, and while this may erode some of the resist thickness, it should not attack the resist laterally. However, the thicker residual-BARC-regions must still be removed, and this may be done with a second, isotropic over-etch-step. The isotropic BARC removal process will also laterally attack the resist feature, and this can give rise to linewidth-control problems. ARC-open may be carried out in a separate etch chamber prior to the oxide, metal, or poly etch, or the ARC may be etched *in-situ* prior to the film being etched. Using a separate chamber approach leads to a larger COO penalty. Hence, dry-etch-tool vendors (such as Lam Research), offer *in-situ* ARC-open processes integrated with the etch process of all film types (Al, W, poly, and oxide). Organic BARCs are typically etched with O_2 chemistries (together with appropriate additives such as Ar, He, and CF_4).

Another issue involving organic BARCs is their (in)compatibility with chemically-amplified (CA) resists. That is, the acidity levels of organic BARCs can negatively interact with CA resists (resulting in acidic or basic by-products being formed during thermal hardening of the resist). These by-products can

create an intermixing layer at the BARC/resist interface. Such intermixing must not be allowed to occur. Researchers have developed organic BARCs for KrF and ArF resists that have chromophore and cross-linkable functional groups in the same copolymer form as the photoresist to avoid this problem.[14] The key difficulty of implementing the solution of closely matching the chemistry of the organic BARC to that of 197-nm resist is that by 2001 there was not yet a mature 197-nm resist. Until a baseline product is chosen, it will be hard to closely match the organic BARC and resist. Only when the industry standardizes to one or two 197-nm resists, will it be possible to apply this solution successfully.

A manufacturing-related challenge of using organic BARCs is caused by the fact that the liquid, spin-on BARC materials coat spinning bowls and coagulate in drain lines. A new material introduced by Nissan Chemicals (DUV44) reportedly uses 70% less catalyst, which results in less spin-bowl coagulation. It retains the optical properties of its predecessors so that there is minimal process variation with respect to depth of focus or exposure latitude.

6.2.2 Inorganic Dielectric BARCs

Inorganic dielectric BARCs are deposited by PECVD. These are typically films of silicon oxynitride (SiON) or silicon nitride (SiN) that function both as index-matching and absorbing BARCs. The thickness, index-of-refraction, and extinction-coefficient of such silicon-oxynitride layers can be tuned to produce an effective BARC layer. The inorganic-CVD-BARCs are thinner (30-50-nm-thick) than the organic BARC layers. These layers also conformally coat the underlying topography as a result of the CVD process. This allows the entire film to have the correct thickness (i.e., the film does not self-planarize). Inorganic BARCs have chemistries similar to the underlying layers, which can make it tricky to remove them after resist-stripping. Thus, some processes leave the inorganic BARC material in place. On the other hand, inorganic BARCs can have a much higher selectivity to the material being etched (e.g., poly-Si or Al) than can organic BARCs (or the photoresist itself). Consequently, the "ARC open" step is not as much of an issue as when an organic BARC is used. In fact, in some processes that use inorganic BARCs, no separate ARC-open step is used. A smaller overetch time is also needed to clear the BARC all over the wafer and large feature sizes than for organic BARC clearing. Dedicated production tools for depositing inorganic BARCs have been developed by such suppliers as Applied Materials and Novellus Systems.

For 193-nm lithography, the composition of the inorganic SiON films must be altered, since most silicon-based inorganic materials are opaque to 193-nm light.

The BARC films for 193-nm lithography will have to change from SiON films (which are silicon-rich films, where 40+% of the film composition is silicon) to films that are much more like an oxide. However, the PECVD process allows such tuning of the composition of silicon, oxygen, and nitrogen of the film over a relatively wide range. By adjusting the CVD process conditions, the BARC film can be tuned to be suitable for *i*-line lithography, 248-nm lithography, or 193-nm lithography.

Inorganic BARCs can cause adverse reactions between the ARC film and the photoresist (akin to the problems that can occur when organic BARCs are used). The culprits are again amine groups (NH_3) that are left on the surface of the SiON films after processing. Deep-UV resists are more sensitive to such interactions than older *i*-line and *g*-line resists. One technique to avoid this interaction is to *in-situ*-deposit a thin oxide on top of the BARC before spinning-on the resist. This oxide becomes an inert barrier between the resist and BARC and has no impact on the properties of the BARC itself. Another way of preventing such interaction is to use an oxygen treatment after the SiON film has been deposited, but before putting down the resist. This effectively removes any amine groups from the BARC surface. The Novellus PEARL® (Plasma-Enhanced Anti-Reflective Layer) includes an *in-situ* nitrous-oxide treatment before the wafer is taken out of the deposition chamber.

6.2.3 PVD-Deposited ARCs

The most common PVD-deposited ARCs are TiN films. TiN films act primarily as absorber-type BARCs, with TiN having a κ-value of 1.1. Their main application is for patterning Al films with *i*-line lithography. Because TiN ARCs are difficult to remove, they are normally left behind and become part of the device (where they may serve in the role of a diffusion barrier). TiN films are reactively sputter-deposited to a thickness of ~80-nm for this purpose. Their use in DUV-lithography is limited by their high-DUV reflectivity (20–30%).

6.2.4 BARCs for Dual-Damascene Applications

When BARCs need to be used with dual-damascene, low-*k*/copper interconnect systems, another set of complex challenges are encountered. The biggest obstacle is patterning transparent dielectric layers (in which the trenches and via holes of the damascene structures must be etched). These dielectric layers may lie over the rough topography of a reflective film such as copper or polysilicon. Unlike the case of applying a BARC directly onto a reflective layer (such as a poly or Al film), in this situation the intermetal dielectric is transparent. Below it are all

kinds of reflective surfaces, at multiple levels. This makes the conventional method of controlling reflection by destructive interference inadequate. That is, a reflection minimized for one local film stack may result in a high reflection for another local film stack. To solve the problem, multilayer ARCs, termed *universal ARLs (UARLs)* are deposited on the dielectric to be patterned (Fig. 6-14). Either two-layers (or sometimes even three layers) are used to create such structures. The bottom layer is opaque to UV light (and reflective), to mask the underlying oxide, while the upper layer cancels out the reflected light by destructive interference. The underlying layer has a higher index of refraction and extinction coefficient so that it is more reflective than absorbing (in the same way that poly and aluminum films behave). Such UARLs are deposited by CVD, in tools from vendors including Novellus Systems and Applied Materials. The UARL formed in the Applied Materials process is termed *Double–DARC.*® Note that the third layer shown in Fig. 6-14 is an oxide film to prevent chemical interaction between the first two BARC layers and the DUV resist.

6.3 PHOTORESIST PROCESSING SYSTEMS

In the early days of IC fabrication the pieces of equipment that performed the individual resist-processing steps were stand-alone tools. As each step was

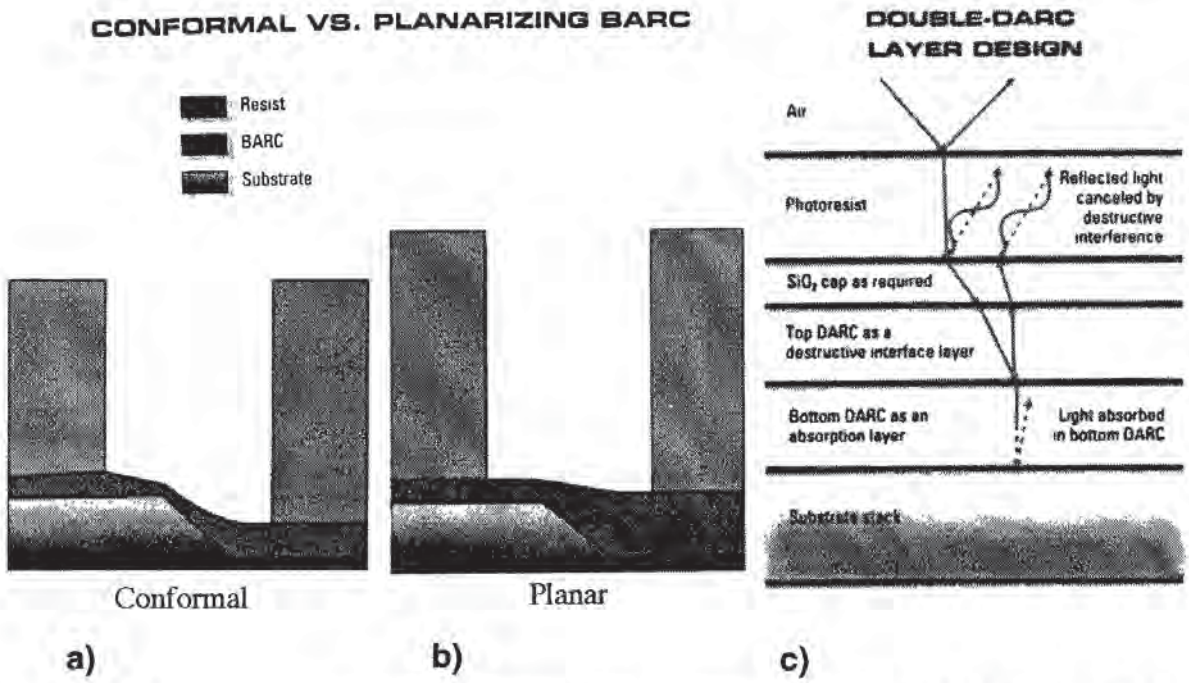


Fig. 6-14 a) Conformal BARC; b) Planar BARC; c) three-layer universal AR layer (UARL).

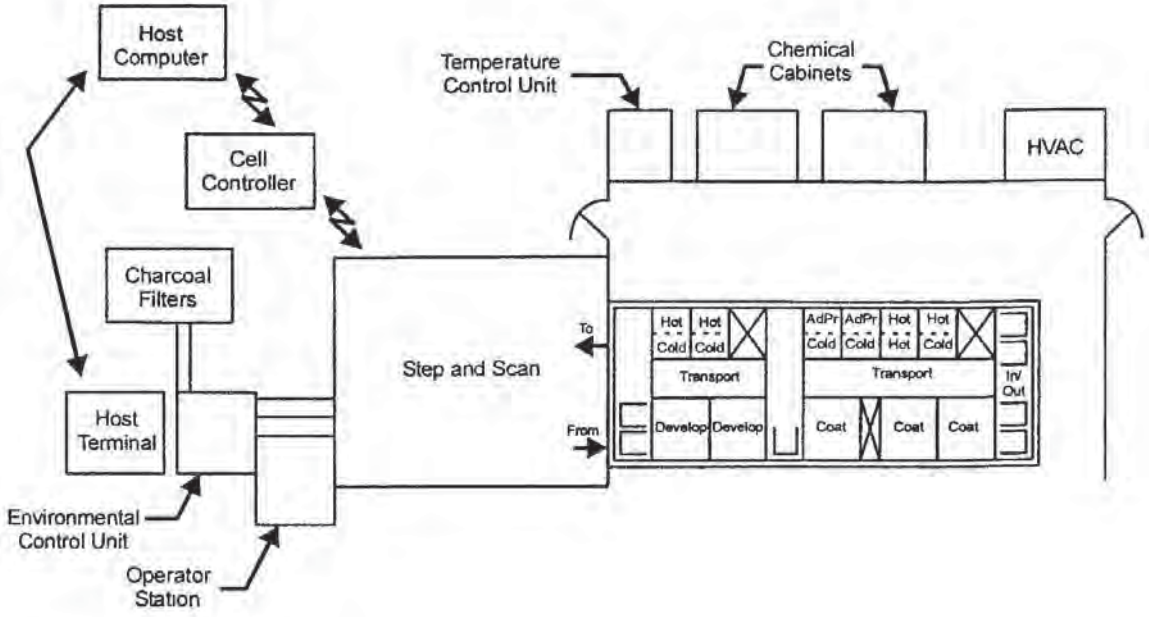


Fig. 6-15 DUV Photo-island.

completed, the wafers would be manually carried from one tool to the next. By about 1990, these tools were integrated into resist-processing systems that automatically moved the wafers from one tool to another without human intervention. Such systems are now configured as either wafer tracks or wafer clusters. In either case, such resist-processing systems are then linked to a stepper to form an integrated stepper-track/photo-island (Fig. 6-15).[15] A cassette of wafers is loaded onto the input station of the photo-island, and 30–60 minutes later they emerge as patterned wafers, ready to be sent to the next process (e.g., ion-implantation or etching).

Wafer tracks originally consisted of two side-by-side lines of resist processing tools (i.e., the prime, spin, and soft-bake modules on the line that fed the stepper, and the PEB, develop, and hard-bake modules on the other). The wafers were transferred sequentially from one module to the next using a belt transport mechanism. In today's *track systems*, the individual tools are still arranged in two lines, but a central robotic transfer shuttle moves the wafers from one tool to another, offering the capability to randomly access the tools (Fig. 6-16). In *wafer clusters* the individual resist-processing tools are arranged in a circular fashion around a central robot that moves the wafers among the tools, again allowing flexible access to each of the process stations.[16,17] The wafer-cluster-system approach was pioneered by Texas Instruments in the mid-1980's and was introduced commercially by FSI in 1991. It has an anthropomorphic, 6-axis

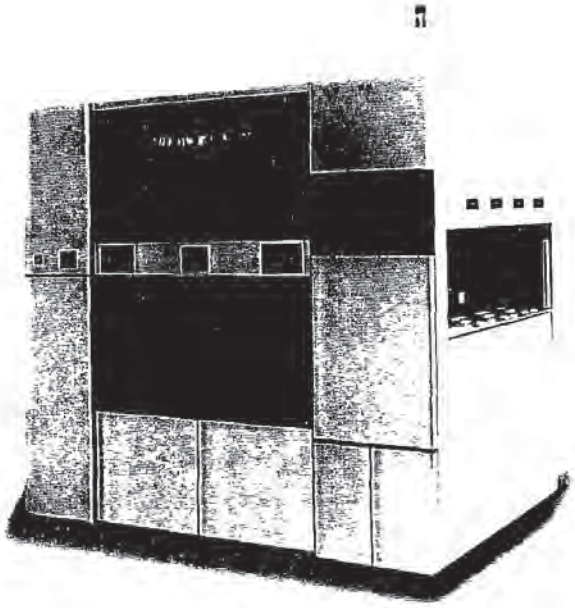


Fig. 6-16 Clean Track ACT platform.® Courtesy of Tokyo Electron America.

wafer-handling system that provides more reach and fewer, larger robots per cluster. SVG's ProCell wafer-cluster system offers a hexagonal platform (Fig. 6-17). Wafers follow a circular path from process module to coating, stepper interface, develop cell, and back to the cassette interface. By scheduling processes with more than 40 wafers simultaneously with varying positions, time requirements and sequencing, the lithography process can be completed in 12 balanced moves. This is done with a patented synchronous scheduling algorithm called *advanced guidance system* (AGS), which controls wafer movement by calculating the optimal wafer path before movement is initiated. System-induced delays are minimized through time management that automatically builds-in scheduled intervals at user-defined non-critical steps, allowing consistent process timing.

The stepper-track (or stepper-cluster) photo-islands are among the most complex systems in a wafer fab. They may consist of up to 20 individual process tools (besides the stepper) that must all work together. At any one moment, the track may be transferring 15-20 wafers through multi-process levels, putting resist on some, developing some, and baking or chilling others. In addition to handling more wafers, the newer lithography processes require the track to perform more steps per wafer. These additional steps include: edge-bead removal,

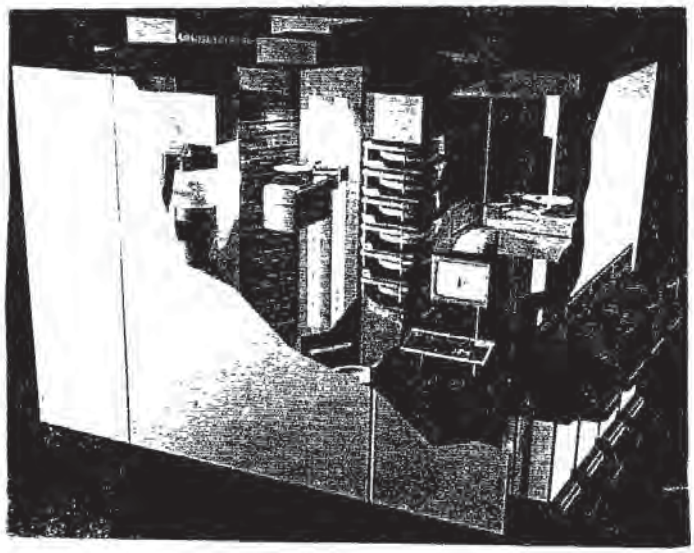


Fig. 6-17 Pro-Cell cluster.® Courtesy of Silicon Valley Group.

applying anti-reflective layers, and processing multi-layer resists. Track systems must also ensure that each wafer (and each lot) has a consistent photoresist coating. The largest manufacturers of such track and cluster systems are Tokyo Electron Ltd., Dainippon Screen, FSI International, and Silicon Valley Group (roughly in that order). We will now describe some of the issues associated with the functioning of photo-islands.[20] They include: capacity, flexibility, reliability, cost of ownership, and environmental/safety features.

The *capacity* of the wafer track is one of its most important characteristics. Advanced wafer stepper/scanners are both fast and expensive ($5M to $10M). To get maximum use from the stepper/scanner, the track must be able to supply wafers as quickly as the stepper can accept them. By 2001, some stepper/scanner throughputs exceeded 100 wafers/hour (200-mm). The wafer track thus requires multiple tools for each process to keep up. For example, as many as six spinners may be needed to support one stepper.

Another critical feature is *flexibility*. The track system must be capable of being modified to meet not only today's, but also tomorrow's needs. It should be possible to add, upgrade, or replace process stations without replacing the entire system. Flexibility is so important because deep-UV resist processing is still evolving, and both hardware and software must be adaptable to "yet to be

determined" queue times, chemistry management, and process steps. However, since 200-mm tracks do not accommodate upgrades to 300-mm wafer sizes, the flexibility show-stopper will be the change to the 300-mm wafer fabrication.

The flexibility capability must also extend to the system software. The software should also be modular, so that it can be reconfigured if needed. The software must also be able to handle multiple process recipes so that an array of processes can be run on the same track (e.g., both *i*-line and DUV resists). The software must meet the SEMI SECS or GEM standards for process-tool-to-host-computer communication. In addition, the operator interface should be well designed so that chances for operator errors are minimized. Ultimately there will be an expert system in the cluster controller that will collect and analyze data from the system using statistical techniques that will determine the overall lithography process status.

The software will also have an impact on throughput. For example, if a new lot cannot be started until the previous lot has finished, the *actual* throughput will be significantly lower than the *steady-state* throughput. To eliminate this problem the software controlling the track must be able to start a new lot while the previous lot is still being processed (a procedure termed *lot cascading*). Hence, it must be possible to start a new lot as soon as the input station is empty. The software must recognize the boundary between the two lots and change the recipe when the first wafer of the new lot arrives at each processing station. Capabilities such as on-line recipe editing and graphical user interfaces also help maximize actual wafer production time.

Next, the system must be highly *reliable*. All wafer track manufacturers specify at least 95% uptime, but even this value means that the system will not be available for 24–36 hours per month. Use of "intelligent redundancy" in subsystems can improve track reliability (for example, by adding an extra hot-plate, or by having qualified standby modules).

The cost of the track system itself, its footprint, and the amount of resist and developer consumption are other factors that impact the *cost-of-ownership*. In the move from 200- to 300-mm wafer tracks, users are demanding that 300-mm tracks be no more than one-third larger than 200-mm systems. Use of vertically stacked modules should allow such system footprint goals to be met. Resist consumption is one of the major costs of operating a semiconductor fab. To control this cost, the dispense volume of resist per wafer has been decreased with time . The design of the resist pump and spinning equipment plays a major role in achieving reduced resist consumption.

The tracks must also be *clean* and *safe to operate.* Both external particles and contaminants must be suppressed, and the track tools and processes cannot generate particles. Point-of-use filtration of resists and developer liquids and advanced wafer transport techniques (that employ gentle handling of wafers) reduces particle generation. Photo-islands are also enclosed in Class-1 environmental chambers. Safety issues involve the control of toxic fumes from resist and developer solutions, and measures to ensure that wafers do not fly off the high-speed spinning chucks.

The environment within the photo-island enclosure must be strictly controlled, including the temperature, relative humidity (RH), and vibration (and for DUV resist processing, the trace quantities of amines).[18,19] The temperature control issue is challenging because the users want an enclosed system in a small footprint. Yet they also require that the temperature throughout this volume (which contains bake modules and motors) also be uniform and tightly controlled. The RH is chosen to be some value in the range from 35–50%, but it must then be maintained at that selected value to ±0.2%. The temperature of the resist and developer chemicals used in the system must also be well controlled.

The goal of the lithographic process is to produce resist features with consistent CDs. In fact, the acceptable CD variation is getting smaller faster than the actual CD size.[20] At 0.18-μm and beyond, good CD control will become the most important challenge of resist processing. Many factors (already mentioned) impact the value of the CD including the resist thickness, the use of an anti-reflection coating, and the exposure and develop processes. For example, in *i*-line lithography, across-wafer and wafer-to-wafer tolerances of 50-Å were acceptable. But sub-0.18-μm processes require < 20-Å ranges. To achieve such tighter film thickness tolerances, the track system must rigorously control cup humidity and temperature, resist temperature, resist dispense rate, and wafer temperature. (Some sub-0.13-μm processes will require ±0.1°C temperature control and humidity within ±0.2%.) The thermal control in baking, as well as the time elapsed between process steps, are becoming additional variables that play critical roles in determining the CD value, especially when DUV-resists are used. In DUV processes, the coating step and the PEB hot-plate-temperature are key, but so is the lag between the time PEB is finished and when the wafer goes on the chill plate. Sequences will have to be timed within 1 or 2 seconds at every step. If the system cannot ensure this transfer timing, CDs will vary from wafer to wafer. (In the past the purpose of the PEB step was to stabilize the image. Now it is the step that basically creates the image. So any variation in the PEB process can significantly impact the CDs for 0.25-μm features and below.)

REFERENCES

1. S. Wolf and R.N. Tauber, *Silicon Processing for the VLSI Era* – Volume 1: Process Technology, Lattice Press, Sunset Beach, CA, 2000, Chaps. 12 & 13.
2. H.J. Levinson, *Principles of Lithography*, SPIE Press, Bellingham WA, 2001.
3. *Microlithography: Science and Technology,* J.R. Sheats & B.W. Smith, Eds., Marcel Dekker, New York, 1998.
4. P.K. Vasudev and P.M. Zeitzoff, "Si-ULSI with a Scaled-Down Future," *IEEE Circuits & Devices*, March 1998, p. 19.
5. H. Ito, "Deep-UV Resists: Evolution and Status," *Solid State Technology*, July 1996, p. 164.
6. C. Garza *et al.,* "Photoresist Materials and Processing," in *Handbook of Semiconductor Manufacturing Technology,* Y. Nishi and R. Doering, Eds., Marcel Dekker, Inc., New York, 2000. Chapter 17, p. 515-528.
7. D. Ruede *et al.,* "The Impact of Airborne Molecular Bases on DUV Photoresists," *Solid State Technology,* August 2001, p. 63.
8. 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.
9. L.J. Peters, "Resist Developers Try to Beat the Clock," *Semiconductor International,* February 2001, p. 77.
10. L.J. Peters, "Resists Join the Sub-Lambda Revolution," *Semiconductor International,* September 1999, p. 68.
11. B. Su *et al.,* "Analyzing and Characterizing Resist Shrinkage," *Solid State Technology,* May 2001, p. 52.
12. "Developments in Materials for 157-nm Photoresists," *Solid State Technology*, October 2001, p. 63.
13. P. Singer, "Anti-Reflective Coatings: A Story of Interfaces," *Semiconductor International*, March 1999, p. 55.
14. L. Peters, "Progress in KrF Antireflective Coatings," *Semiconductor International,* May 2000, p. 25.
15. P. Burggraaf, "Stepper-Track Integration: Solving the Problems," *Semiconductor International,* August 1992, p. 30.
16. R. DeJule, "DUV Cluster Tools Go Mainstream," *Semiconductor International,* August 1999, p. 44.
17. M. Krishna *et al.,* "Improve Yields and Enhance CDs. With Integrated DUV Resist Track," *Solid State Technology,* October 2000, p. 135.
18. B. Demandante *et al.,* "Protect DUV Processes with Real-Time Molecular Monitoring," *Semiconductor International,* September 2000, p. 115.

19. L. Peters, "Maintaining a Clean Environment in DUV Lithography Bays," *Semiconductor International,* May 1999, p. 52.
20. A.E. Braun, "Track Systems Meet Throughput and Productivity Challenges," *Semiconductor International*, February 1998, p. 63.
21. L. Peters, "Readying Resists for the 90-nm Node," *Semiconductor International,* February 2002, p. 63.

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.

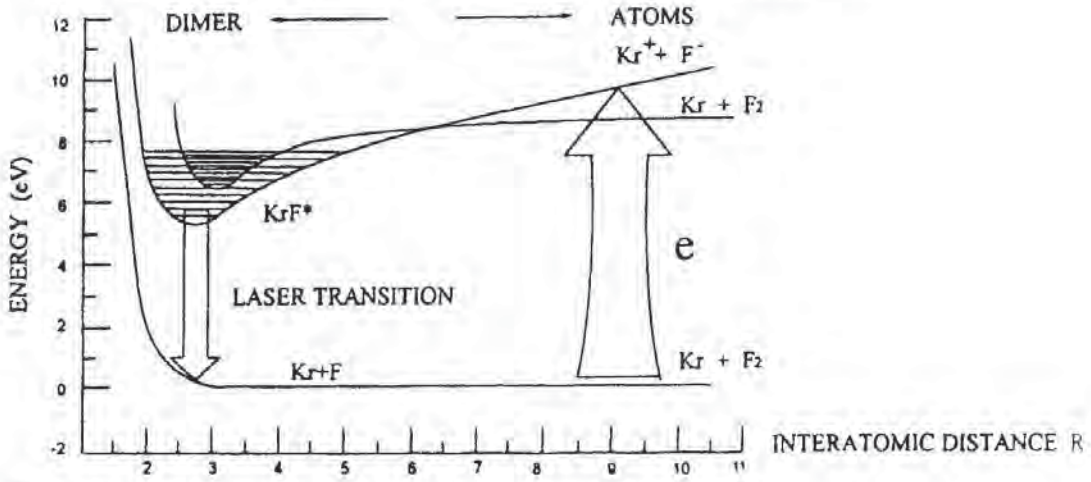


Fig. 7-1 Energy diagram for a KrF excimer laser. KrF^* is formed via two reaction channels. It decays to the ground state via dissociation into Kr and F while emitting a photon at 248-nm.

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

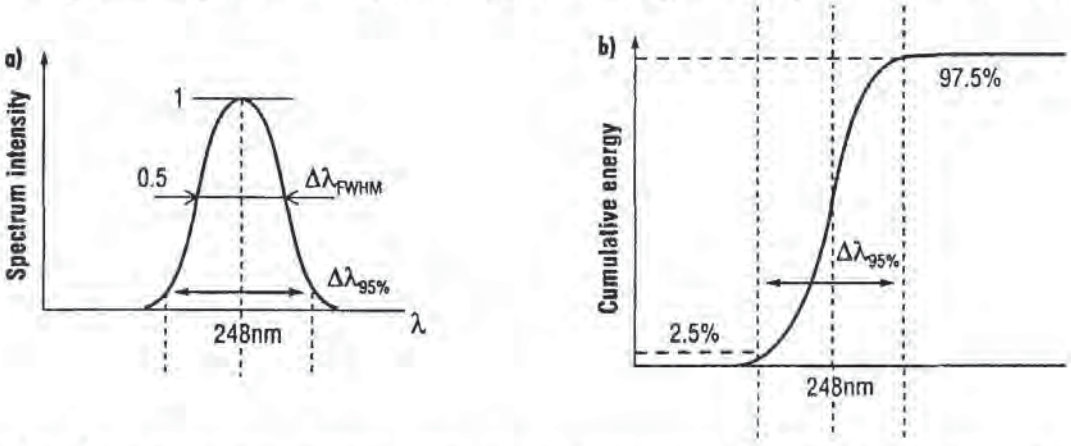


Fig. 7-2 a) Definition of "full width at half maximum," FWHM; b) Definition of 95% integral pulse bandwidth.

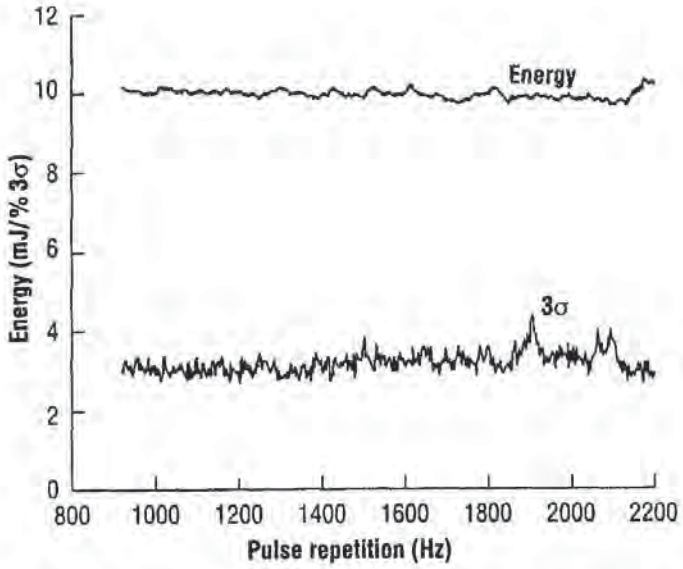


Fig. 7-3 Pulse energy and stability vs. repetition rate for a KrF laser running in the constant charge voltage mode.

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

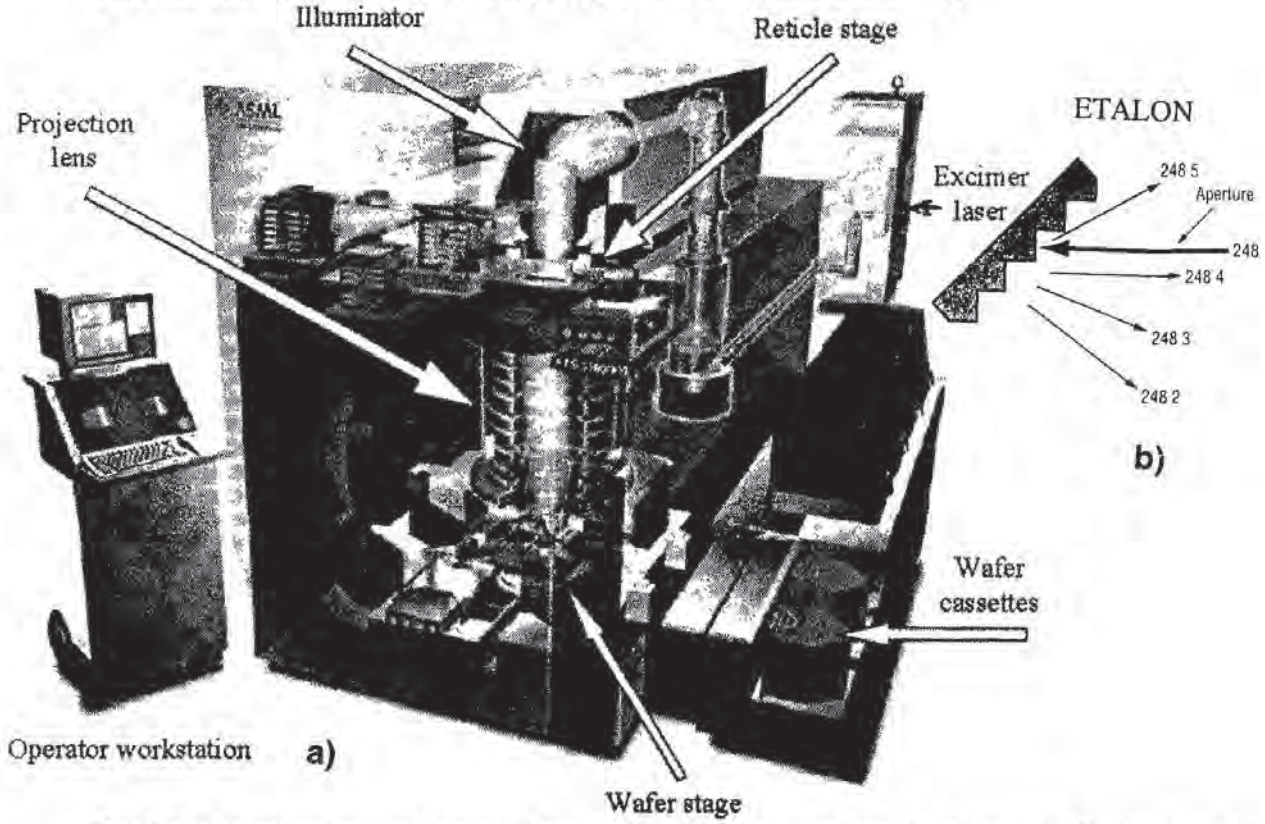


Fig. 7-4 a) Configuration of excimer laser light source and stepper (Courtesy ASML); b) Etalon

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 2×10^9 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^{TM} 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

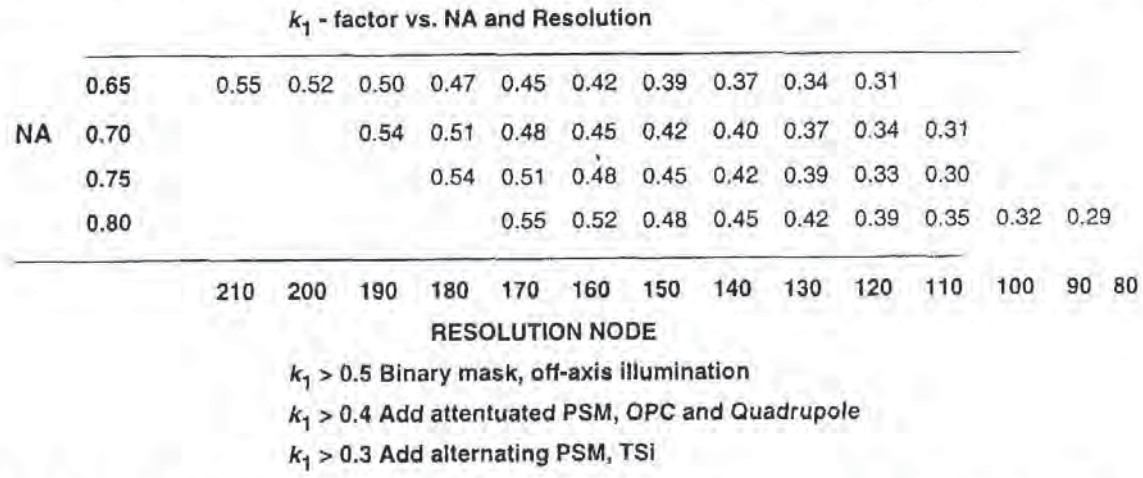


Fig. 7-5 The 248-nm lithography tools have seen vast improvements in lens aberrations in recent years. (Courtesy of ASML)

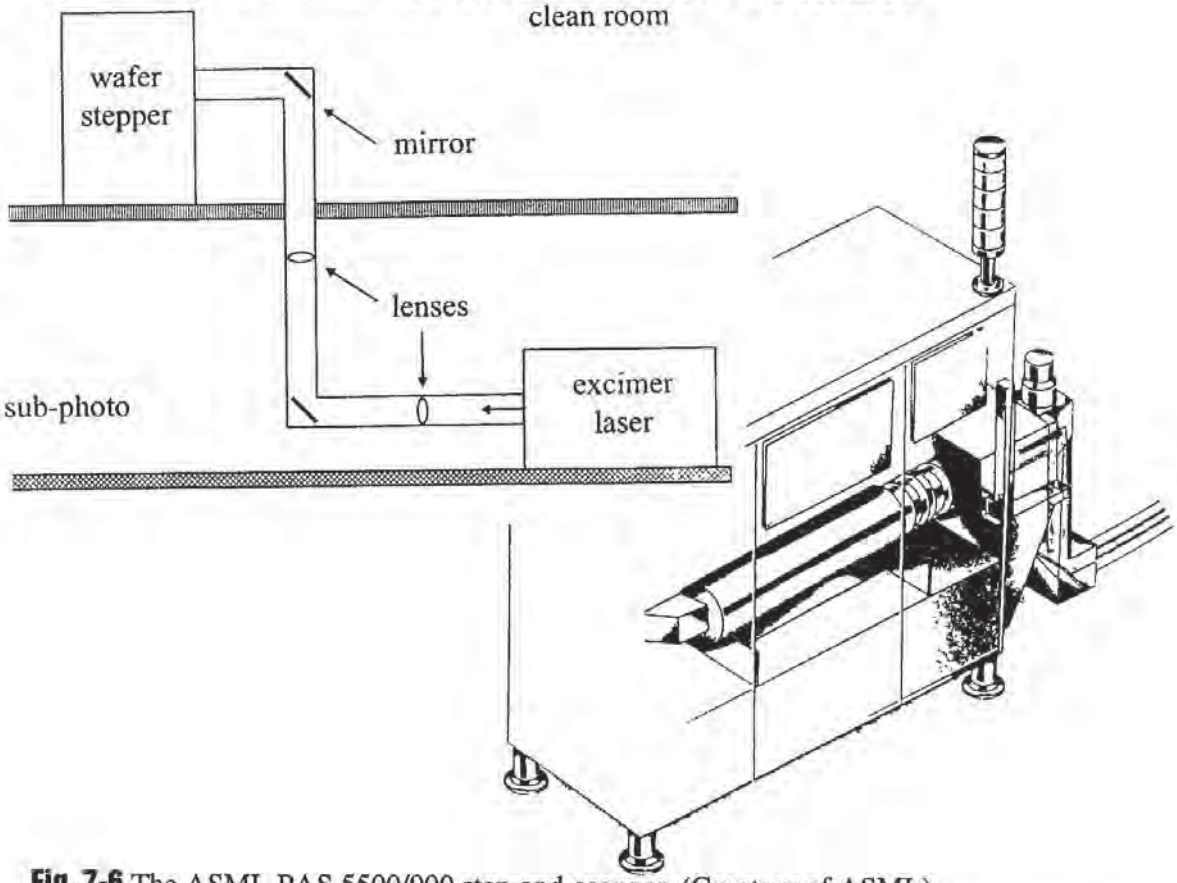


Fig. 7-6 The ASML PAS 5500/900 step-and-scanner. (Courtesy of ASML)

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)

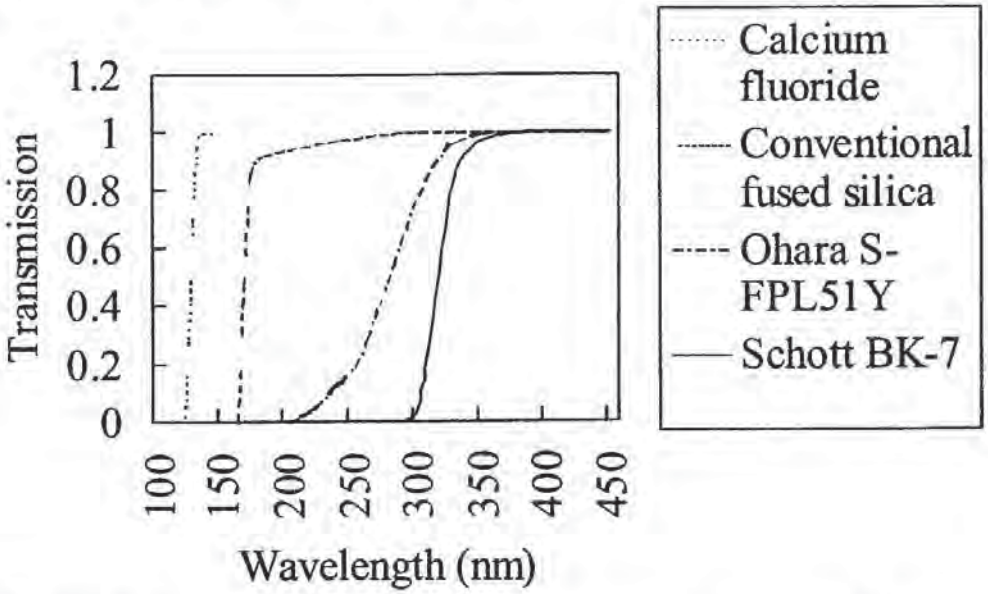


Fig. 7-8 Transmission of 10-mm-thick samples of materials used in lenses of wafer steppers.

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

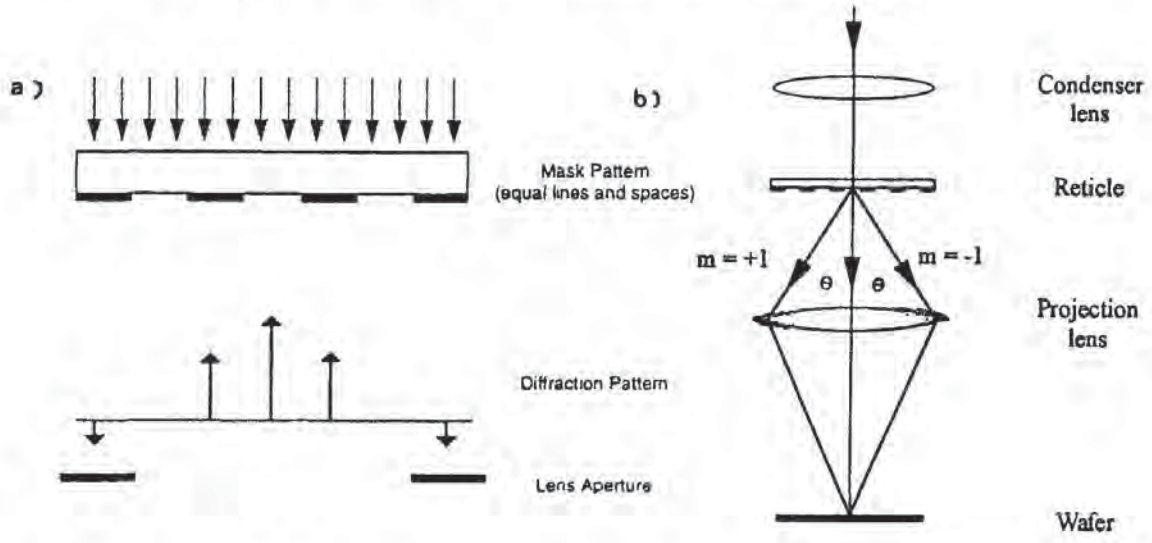


Fig. 7-9 Principle of on-axis illumination: a) Light illuminating the mask from an on-axis direction, and the diffraction pattern from an on-axis coherently illuminated grating mask. The pattern is centered on the entrance pupil of the objective lens. b) Aperture with on-axis illumination showing the 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

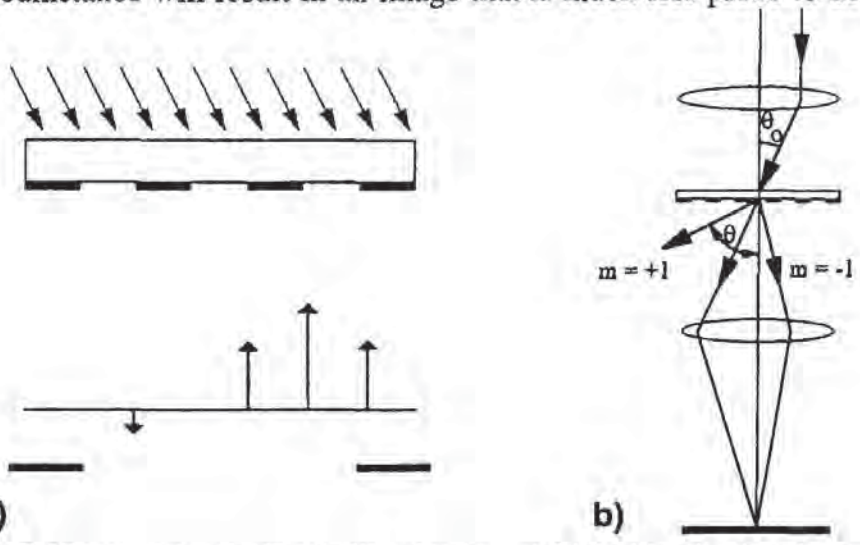


Fig. 7-10 Principle of off-axis illumination: a) Light illuminating the mask from an off-axis direction, and the diffraction pattern from an off-axis coherently illuminated grating mask is no longer centered on the entrance pupil of the objective lens, but is instead shifted. b) Aperture with one off-center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

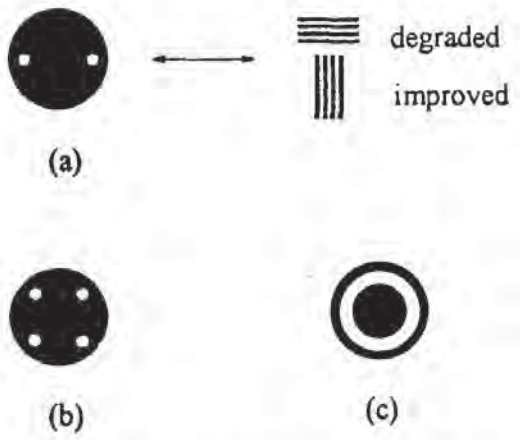


Fig. 7-11 Different types of off-axis illumination apertures. The aperture with holes at the edges or corners - (a) and (b) – will improve lines oriented in one direction (see a1), but will degrade imaging of lines in the orthogonal direction. (c) Annular off-axis illumination aperture.[23]

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

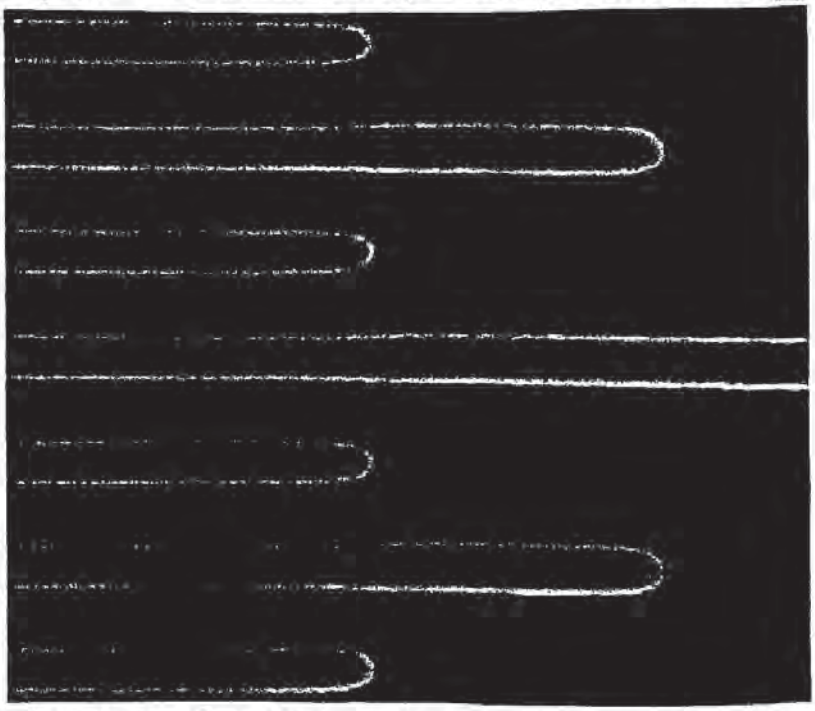


Fig. 7-12 Optical proximity effect. All pattern features are the same size on the reticle, but the final image is larger for the isolated line extension than for the densely packed lines.

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

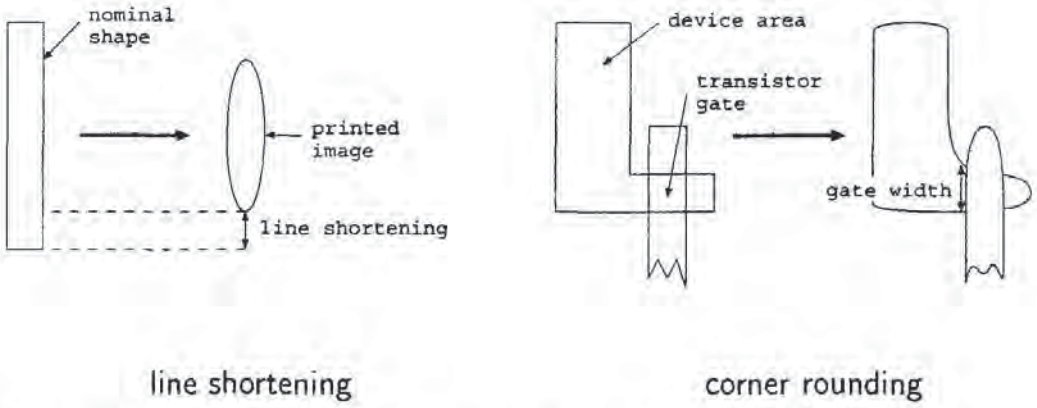


Fig. 7-13 Example of line shortening due to image distortion.

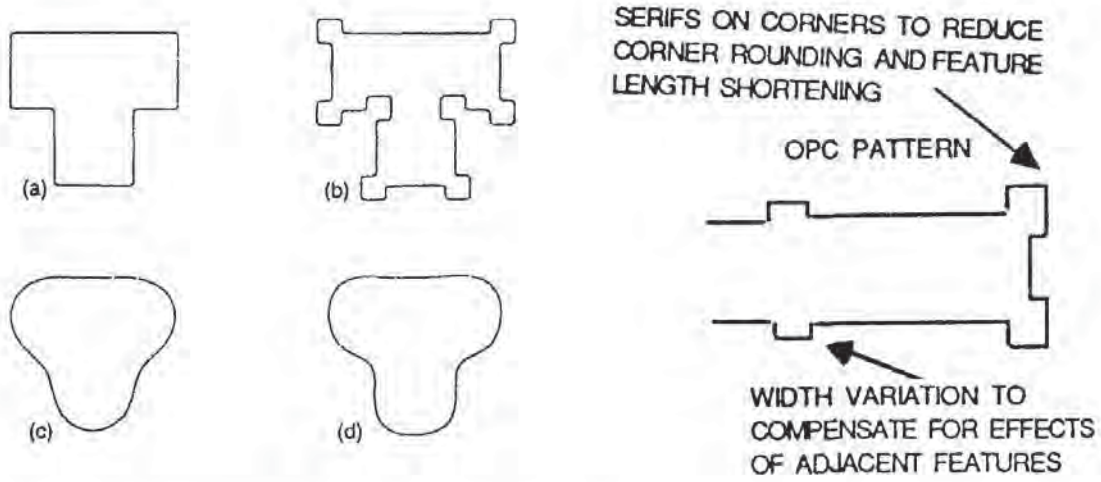


Fig. 7-14 A T-shaped feature (a) with dimensions near the resolution limit of the projection lens is printed as a rather featureless blob (c) on the wafer. Addition of serifs (b) brings the printed image (d) closer to the shape originally designed. Reprinted by permission of Marcel Dekker. (e) Adjustments 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

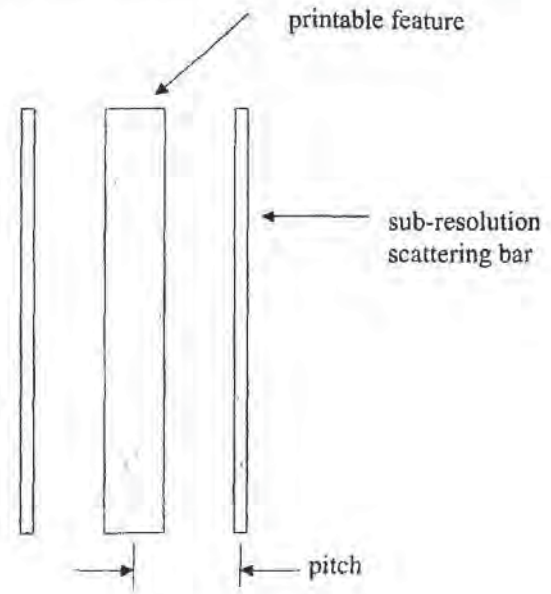


Fig. 7-15 Sub-resolution scattering bars applied to isolated features.

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

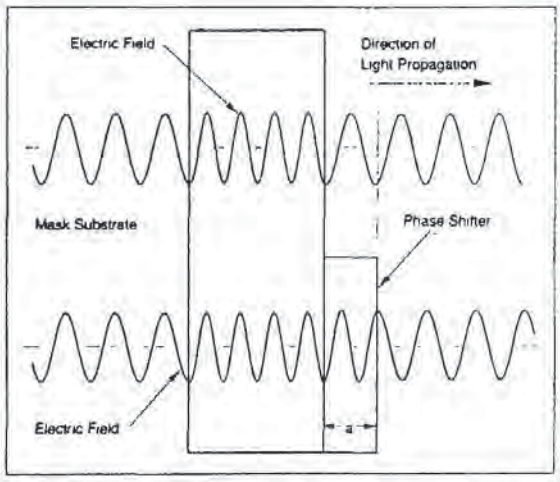


Fig. 7-16 The phase-shifting layer introduces a difference in optical-path-length. Here the difference is $\lambda/2$.[30] © IEEE (1993)

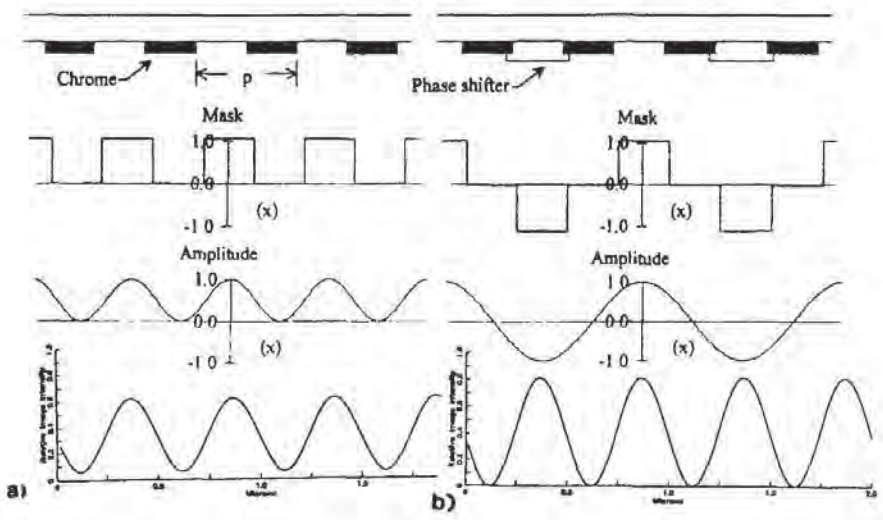


Fig. 7-17 Schematic of (a) a conventional binary mask (b) an alternating phase shift mask, (ALT PSM). The mask electric field, amplitude, and image intensity is shown for each type of mask.[28] Reprinted with permission of Marcel Dekker.

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-

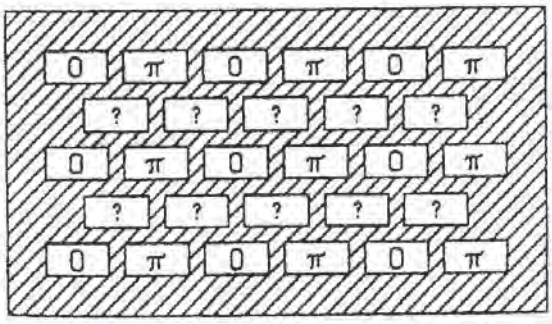


Fig. 7-18 Limitation of ALT-PSMs. There is no way to assign phases so that a 180°-phase difference occurs between adjacent transparent features In this example, the phases have been assigned to the odd-numbered rows, but there is no way to assign phases consistently to the even rows.[36] Reprinted with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

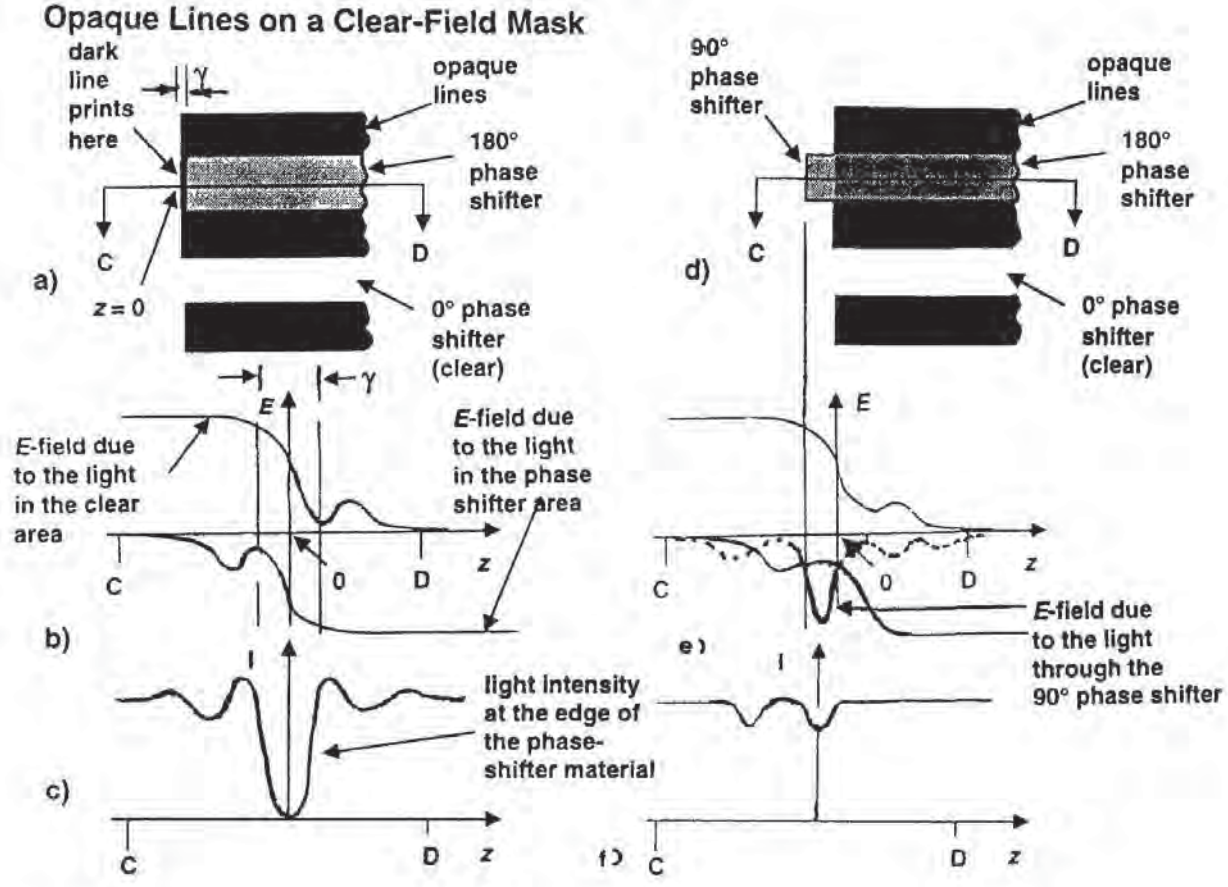


Fig. 7-19 Problem at the edge of a phase-shifter at a clear boundary: (a) ALT-PSM used on lines and spaces pattern with a clear field. Dark lines will print at the boundaries between the edge of the phase shifter and non-shifted clear areas; (b) *E*-fields along the *z*-axis between C and D for light 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-

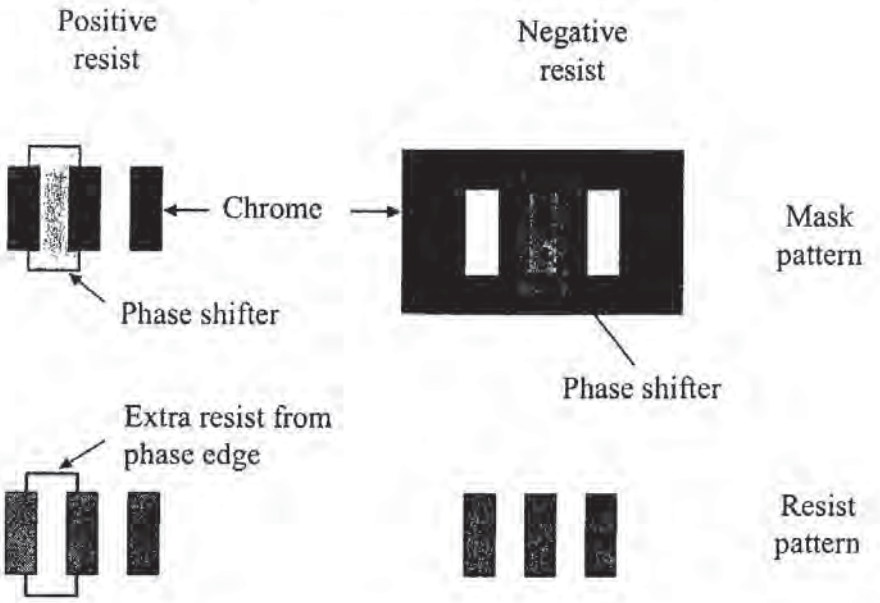


Fig. 7-20 On the left, extra resist printed by the phase edge, for positive photoresist. The right side of the figure shows what occurs with negative resist.

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

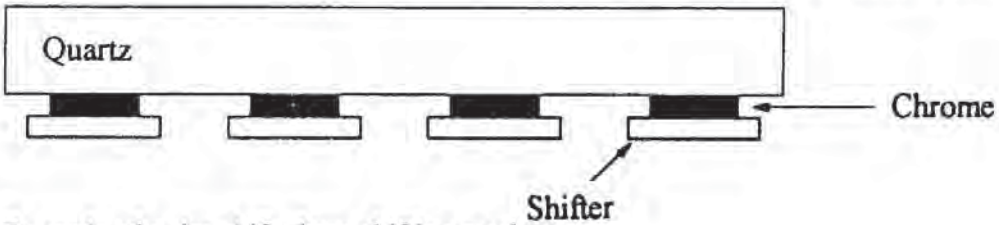


Fig. 7-21 Example of a rim-shift phase-shifting mask.

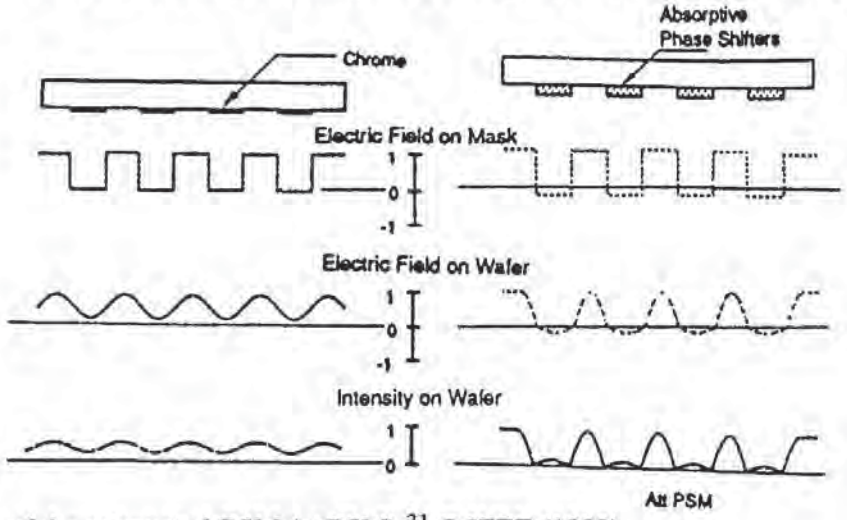


Fig. 7-22 Principle of the attenuated-PSM (attPSM).[31] © IEEE (1993)

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.

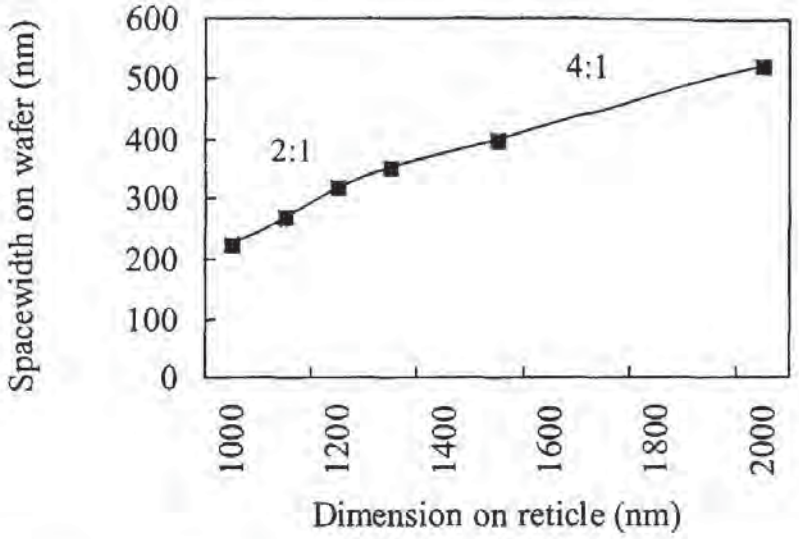


Fig. 7-23 Transfer of features from a reticle to the wafer. The features are reduced in size by 4x for large feature size, but the reduction changes for features near the resolution limit of the lens.

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.

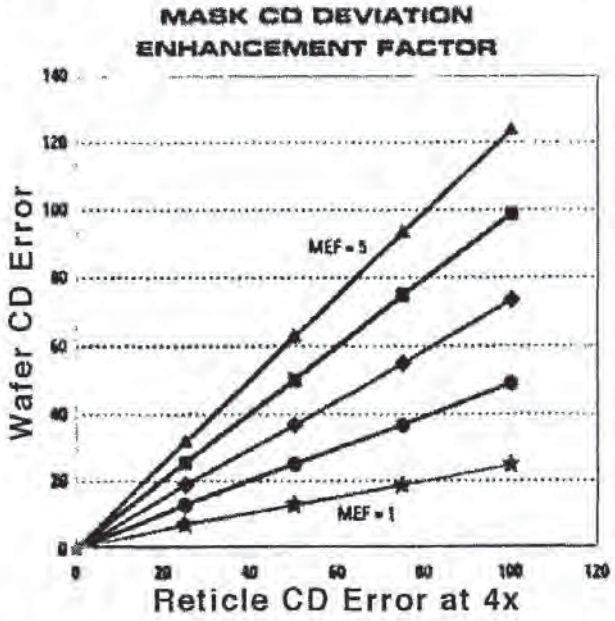


Fig. 7-24 For a MEF >1, a small CD-error on a 4X mask can transfer a larger magnitude of error onto the wafer. (Courtesy KLA-Tencor)

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 $\pm$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

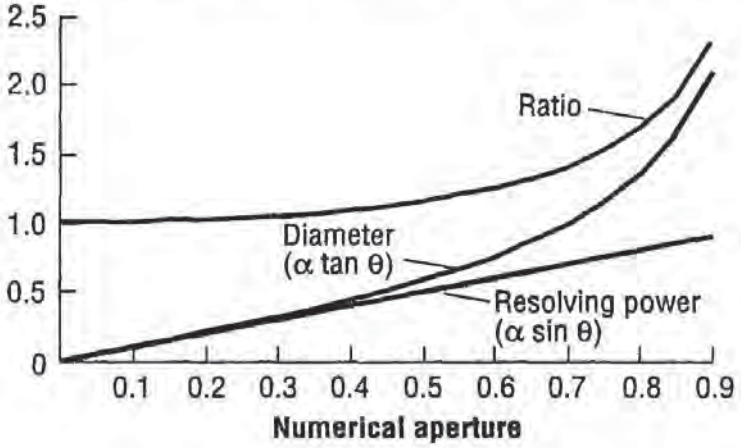


Fig. 7-25 Diminishing returns of higher NA for higher resolution. For lens with an NA greater than 0.80 NA, each incremental increase in lens-resolving power requires disproportionately larger increases in lens diameter. The ratio represents the difference of lens diameter to resolving power at a given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

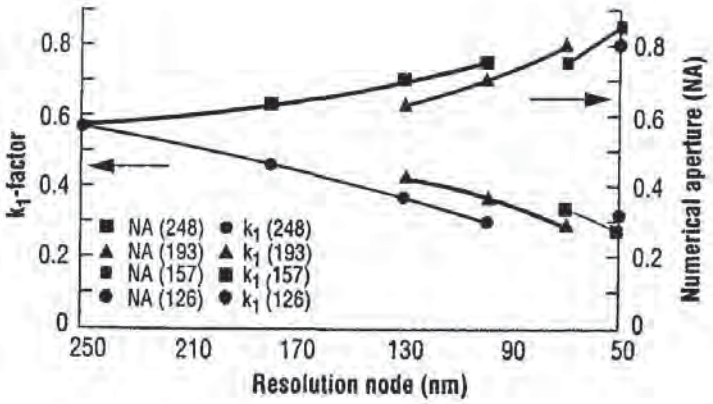


Fig. 7-26 Wavelength, NA, and k_1 for 250 nm to 50 nm resolution nodes. (Courtesy ASML)

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

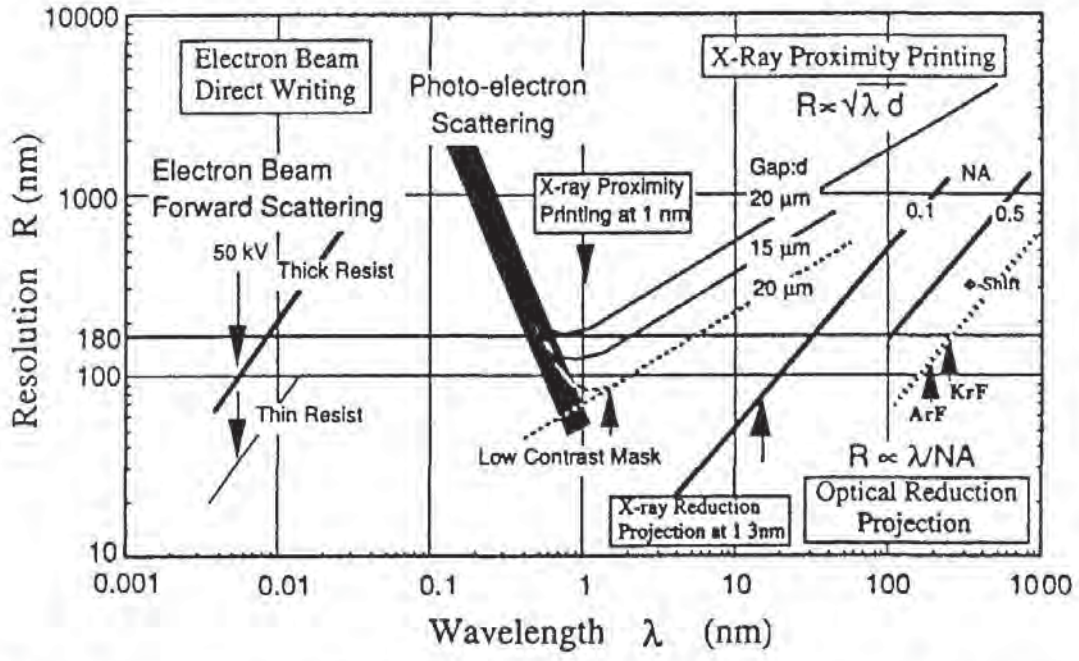


Fig. 7-27 Relationship between resolution and exposure wave length for various microlithographic technologies.[42] © 1996 (IEEE).

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

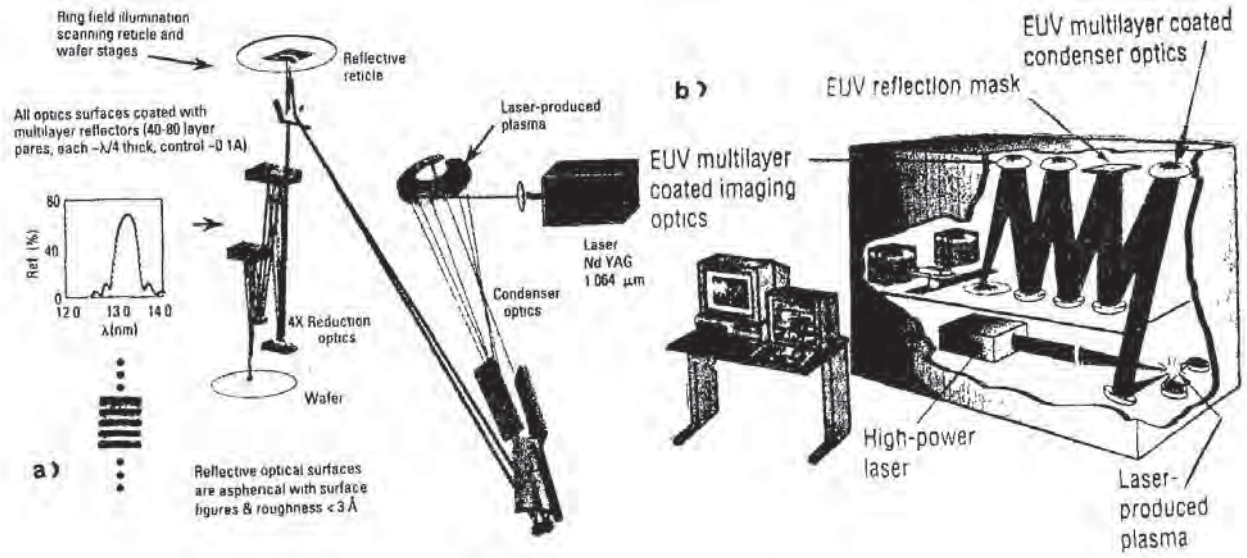


Fig. 7-28 (a) Schematic drawing of the optics of an EUV system. (b) An artists drawing of an EUV lithography system. A laser generates EUV radiation by exciting a laser produced plasma. This radiation is collected by a multilayer condenser system. The radiation illuminates a multilayer coated reflection 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

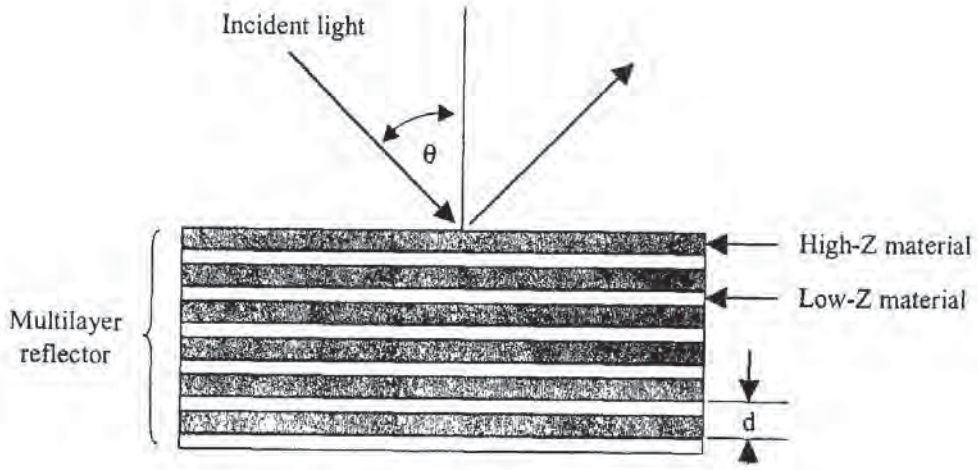


Fig. 7-29 Principle of a multilayer reflector for EUV radiation.

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*

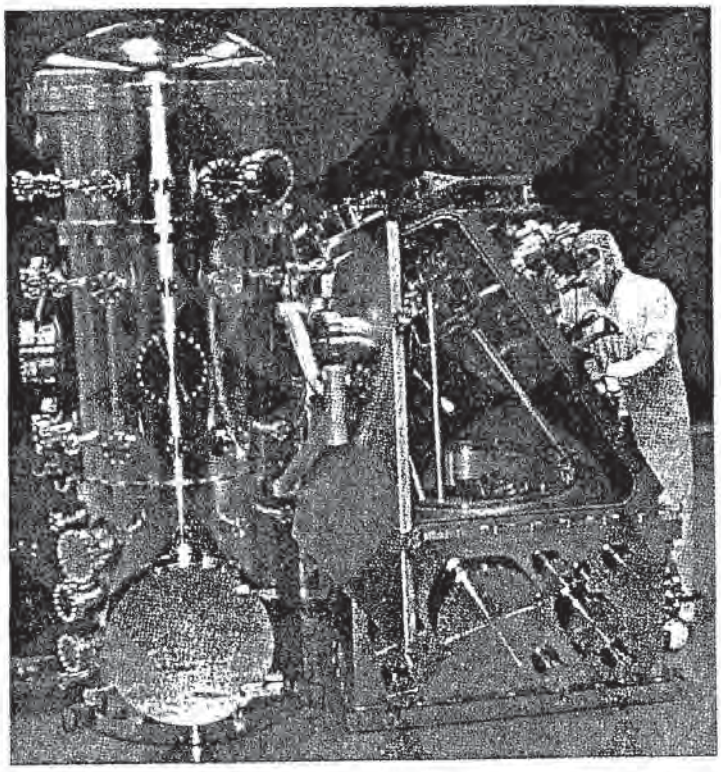


Fig. 7-30 A worker inspects the Engineering Test Stand, a full-scale prototype EUV lithography system – as assembled at Sandia National Laboratories. (Courtesy Sandia Labs).

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

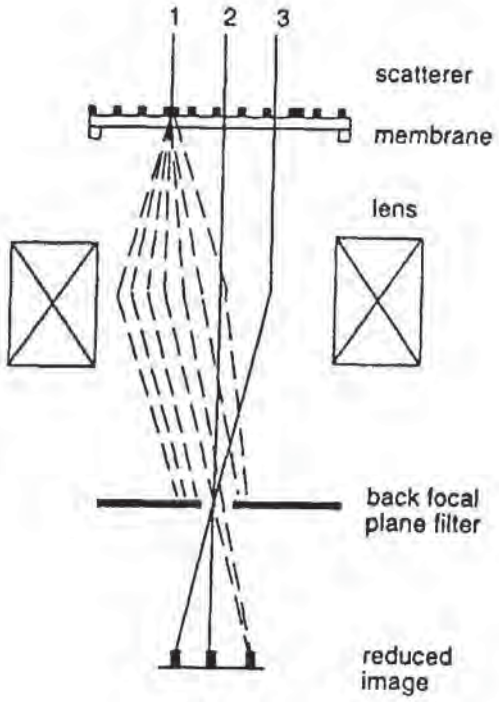


Fig. 7-31 Schematic of the SCALPEL technique. Electrons (1) that hit the scatterer (the metal patterns on the mask) are strongly scattered, and most are filtered out by the aperture. Electrons traveling through the silicon nitride membrane (2, 3) are only weakly scattered and pass through the aperture. 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

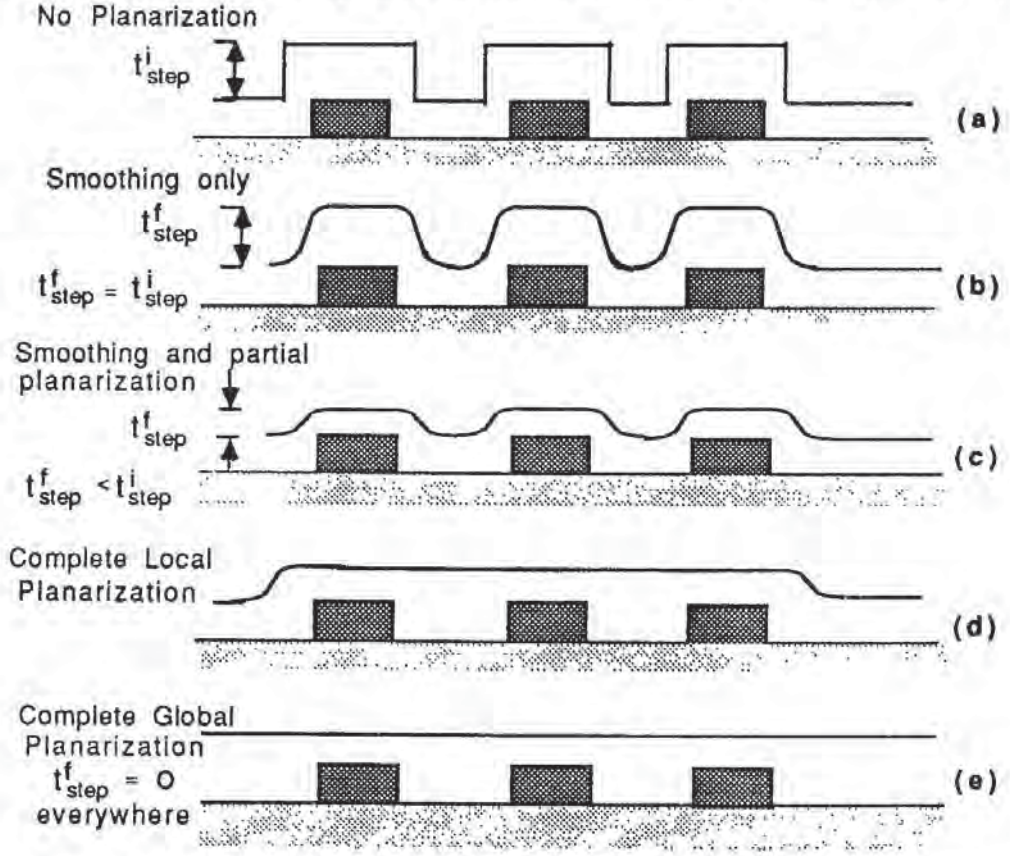


Fig. 8-1 (a) through (e) - Degrees of dielectric planarization; (f) SEM of 6-metal-level interconnect structure, with complete global-planarization realized at each level.

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]

$$\beta = 1 - (t^{f}_{step}/t^{i}_{step}) \qquad \textbf{(8.1)}$$

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:

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

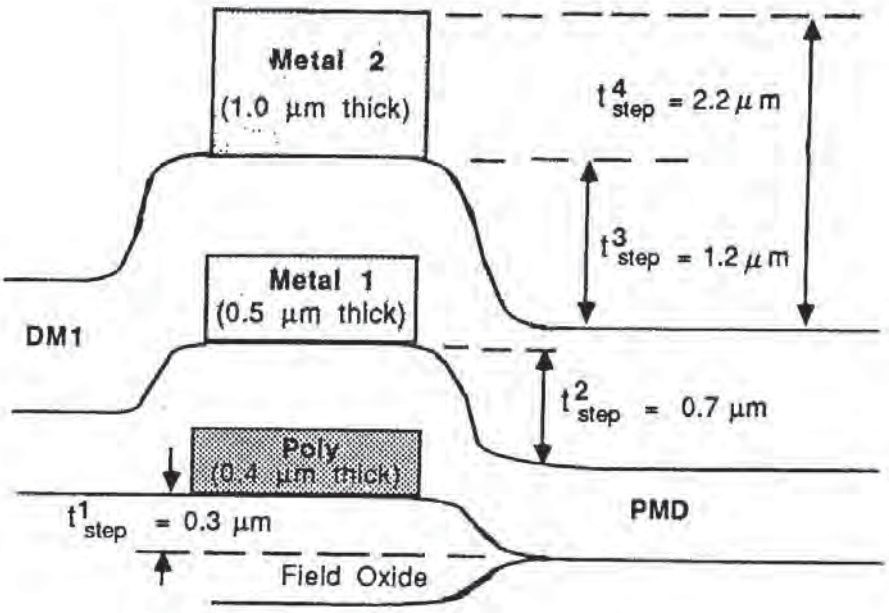


Fig. 8-2 As the number of levels in an interconnect technology is increased, the stacking of additional layers on top of one another produces steps of progressively greater heights.

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

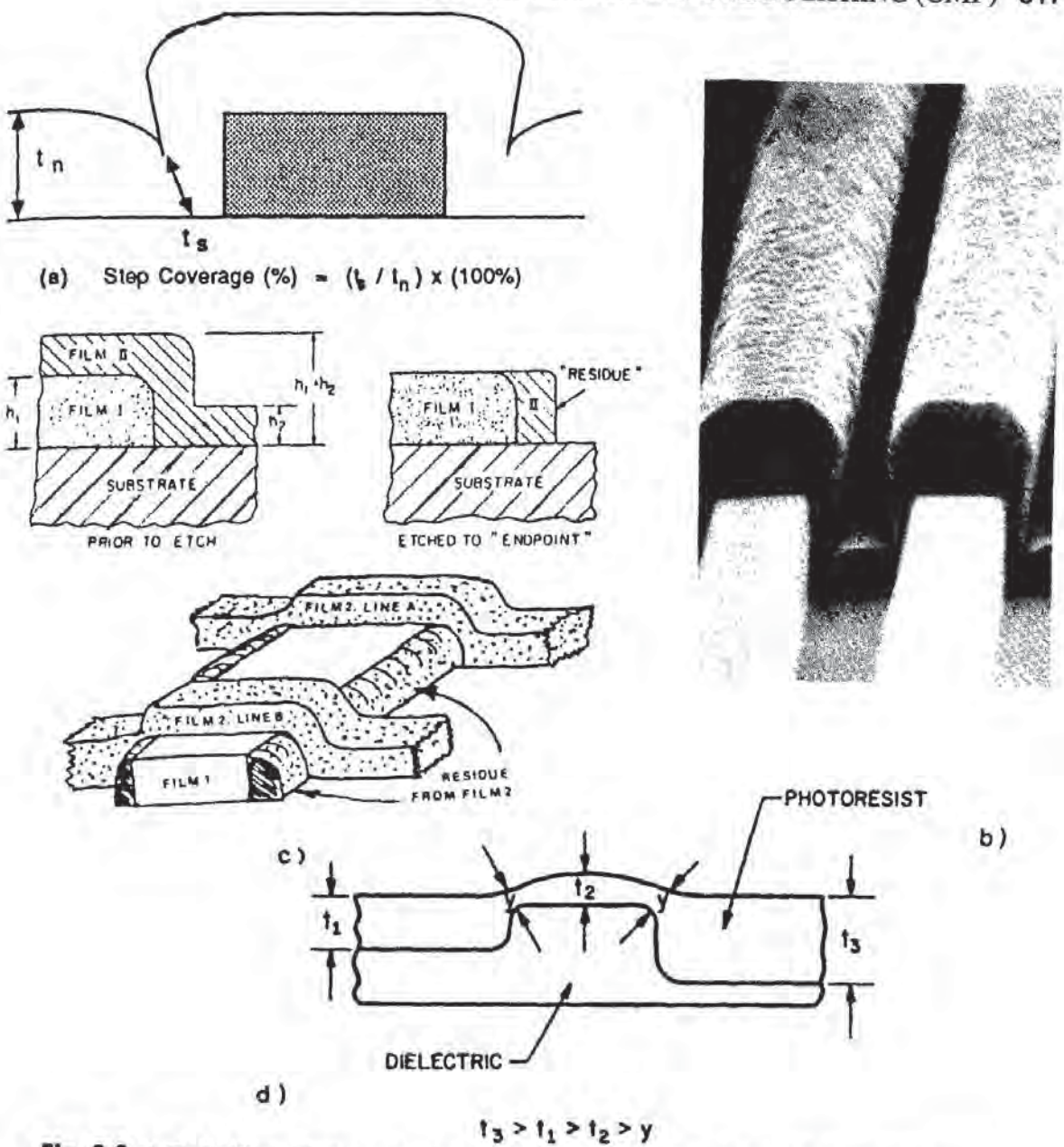


Fig. 8-3 a) Definition of step-coverage. b) SEM micrograph showing poor step-coverage. c) Metal stringers remain behind at the foot or sides of a steep step when anisotropic-etching is used. d) Problem of resist-thinning as film crosses underlying steps.

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

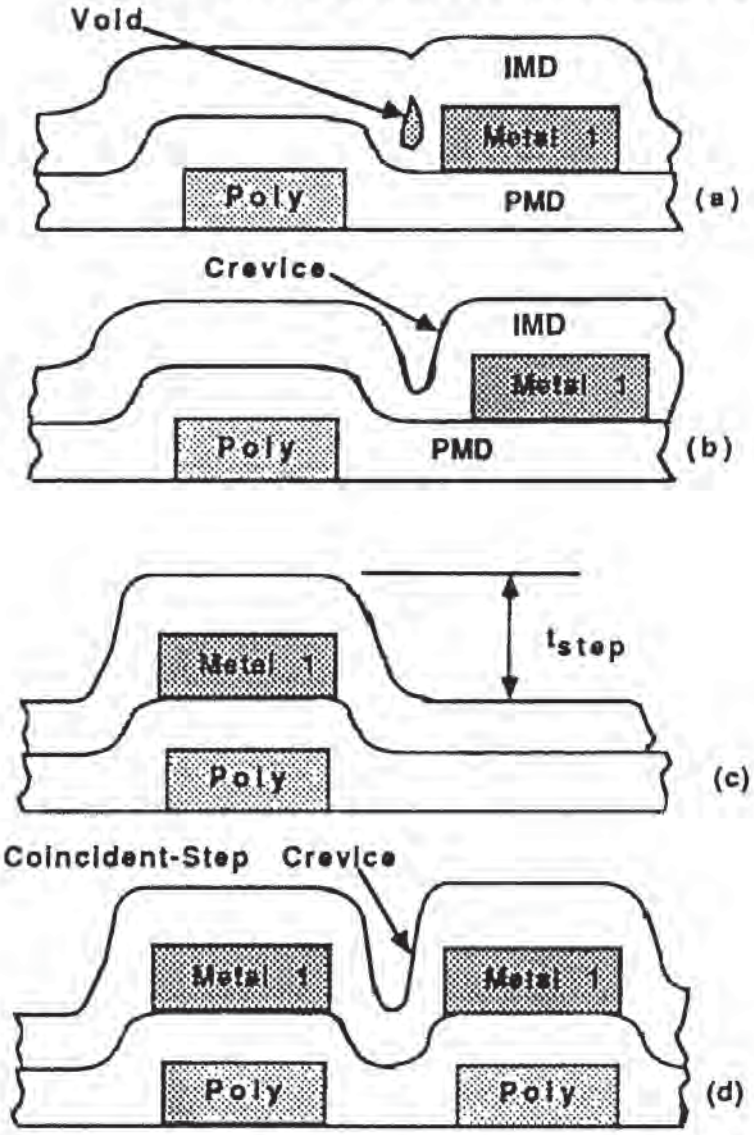


Fig. 8-4 Conditions associated with the relative spacing of polysilicon and Metal-1 that lead to layout design rules of these layers.

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-

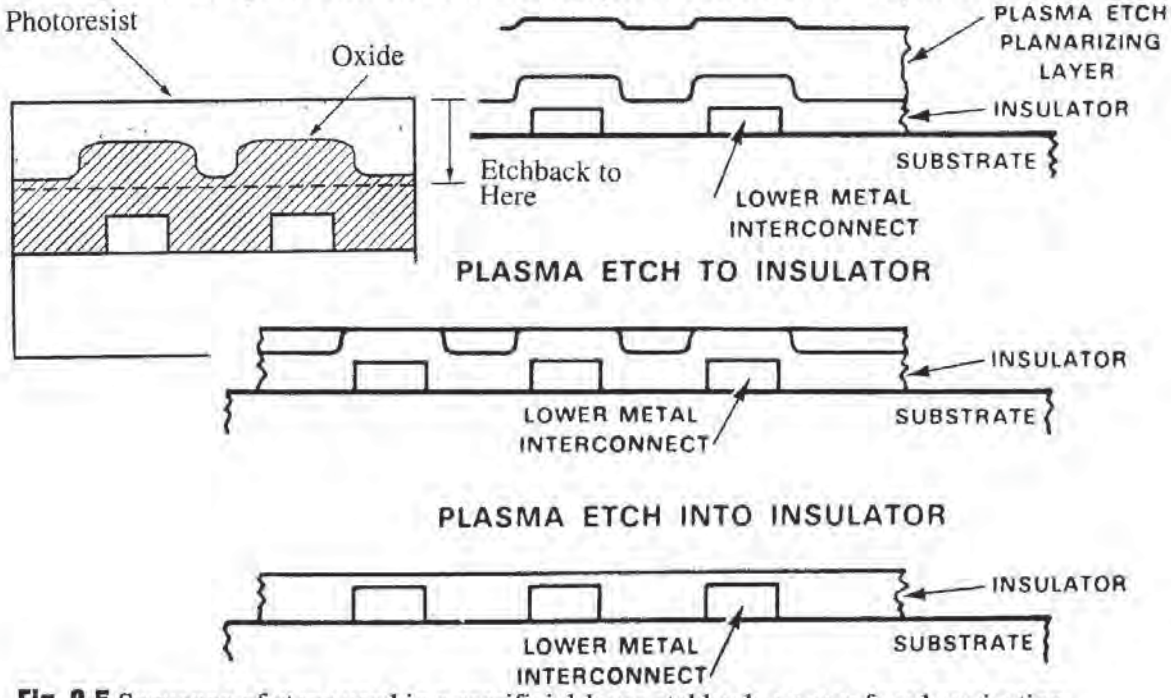


Fig. 8-5 Sequence of steps used in a sacrificial-layer etchback process for planarization.

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

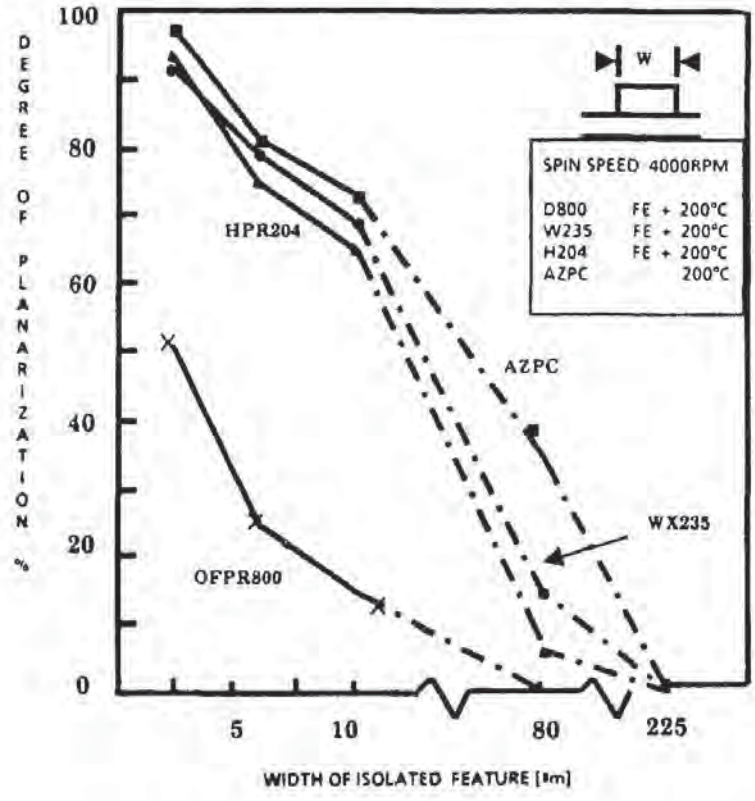


Fig. 8-6 Degree of planarization vs. various feature size for different polymers.[2] (© 1987 IEEE)

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).

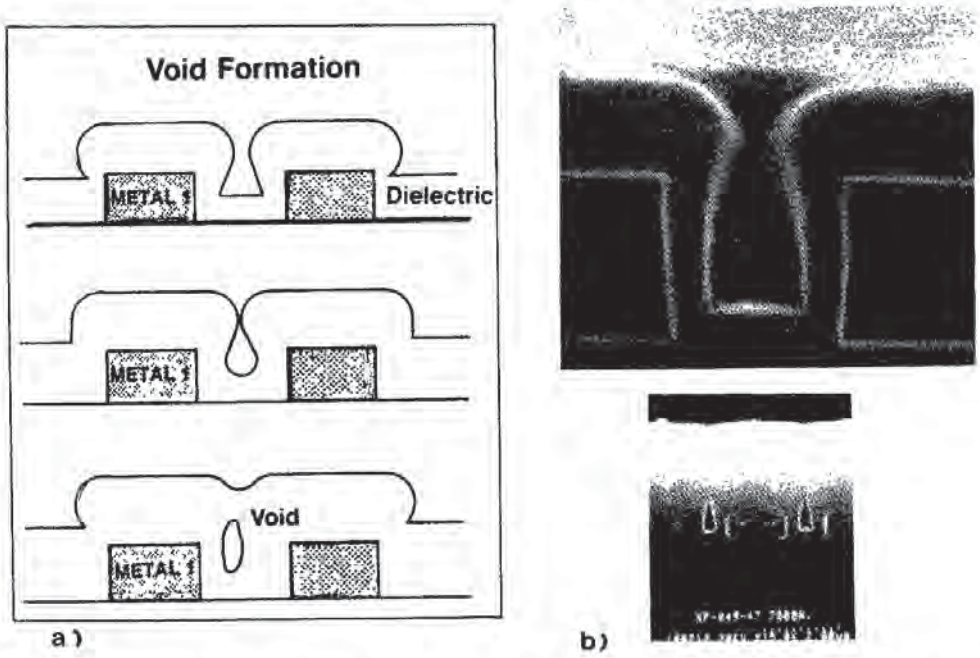


Fig. 8-7 (a) Diagram of void-formation due to cusping of the dielectric. (b) SEM of voids left between 1.5-μm-spaced Metal-1 lines in a 3.0-μm-thick PMD PECVD oxide.

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,

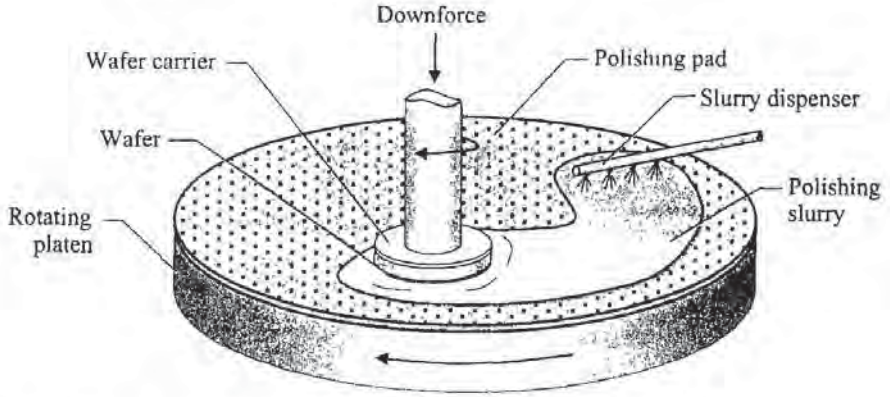


Fig. 8-8 Schematic of a rotary-table chemical-mechanical-polishing system.

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:

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

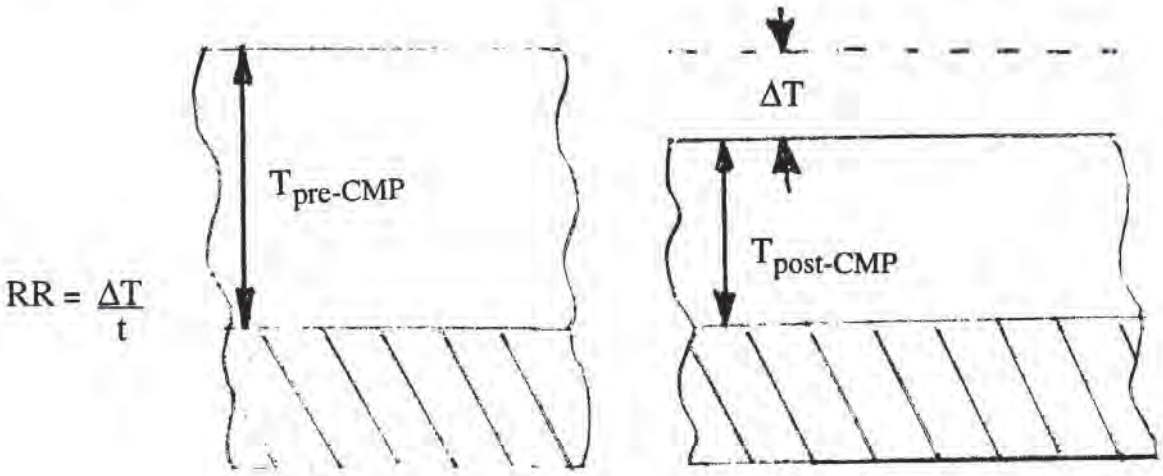


Fig. 8-9 Definition of *removal-rate* (RR).

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, $\bar{x}_{WID}$:

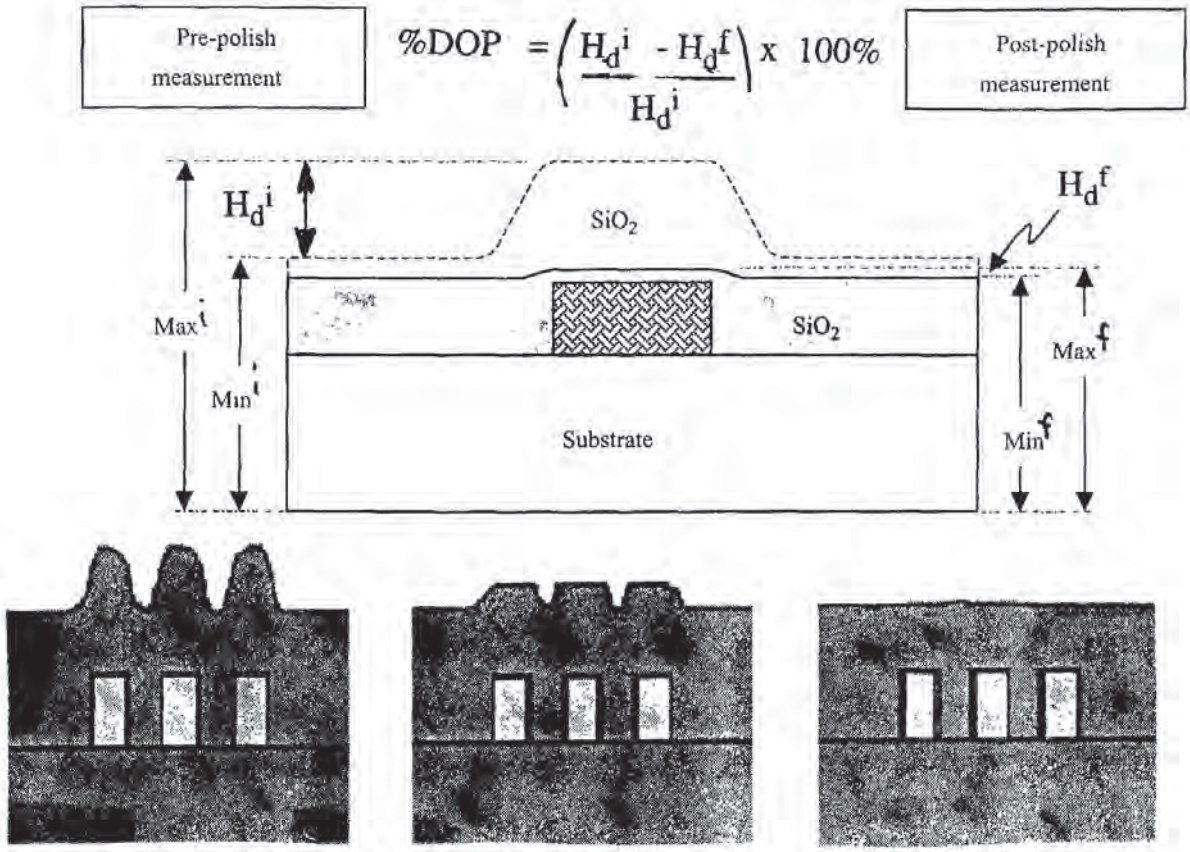


Fig. 8-10 Definition of *degree of planarization* (DOP).

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

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.

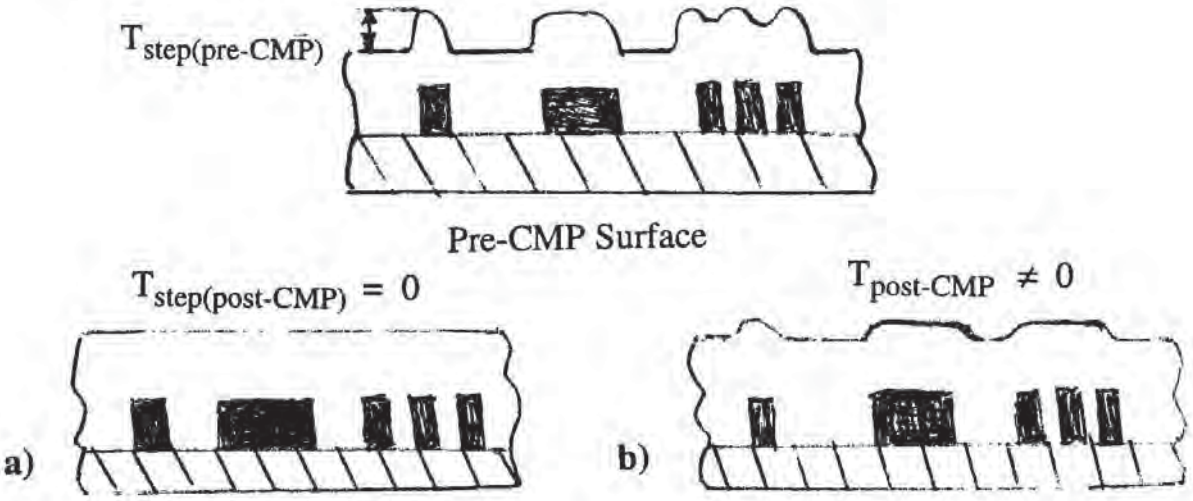


Fig. 8-11 a) Surface on which complete global-planarization has been achieved; b) Surface on which complete global-planarization has not been achieved.

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.

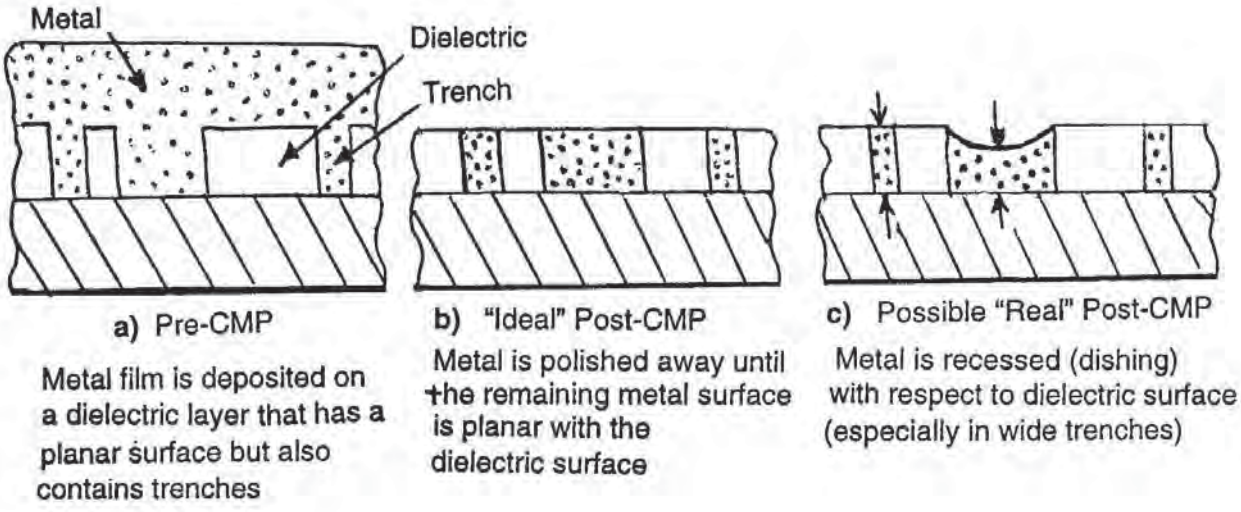


Fig. 8-12 Effect #1 that causes *within-die non-uniformity* (WIDNU).

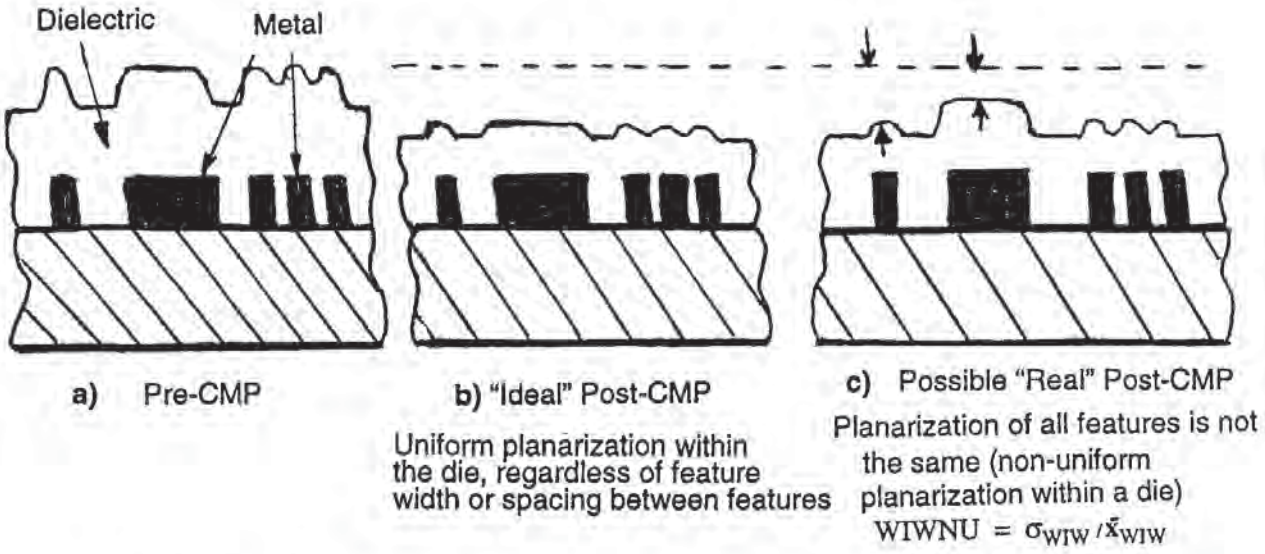


Fig. 8-13 Effect #2 that causes *within-die non-uniformity*, WIDNU (dishing).

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:

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

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.

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

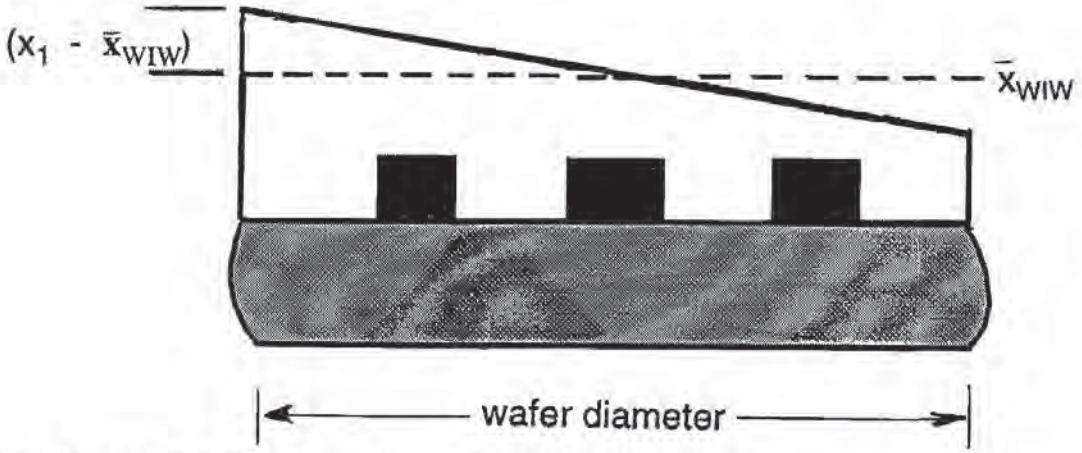


Fig. 8-14 Definition of *within-wafer non-uniformity* (WIWNU).

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

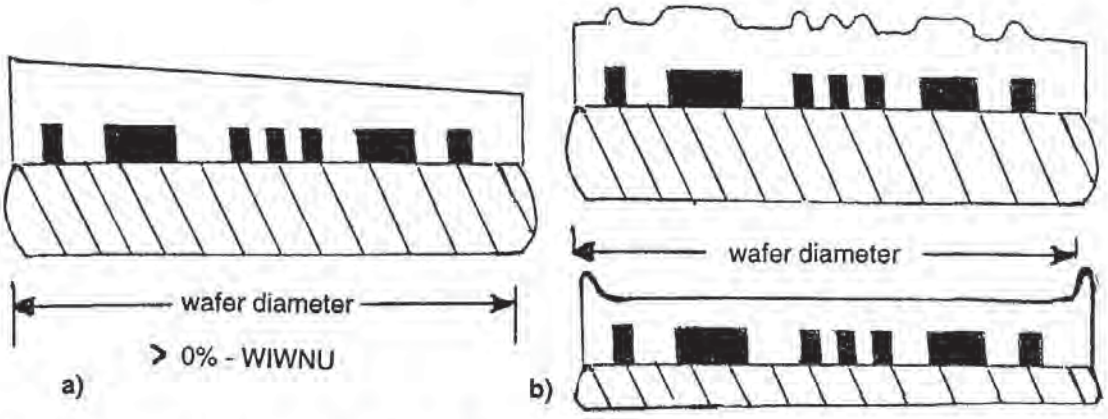


Fig. 8-15 a) Example of one type of WIWNU; b) Example of a second type of WIWNU caused by the *edge-bead-effect*.

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

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

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:

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

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-

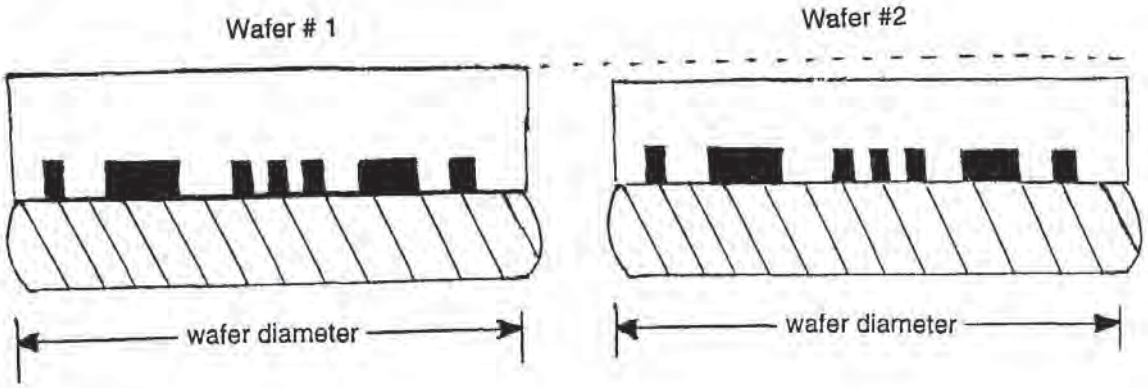


Fig. 8-16 Definition of *wafer-to-wafer non-uniformity* (WTWNU).

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-

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

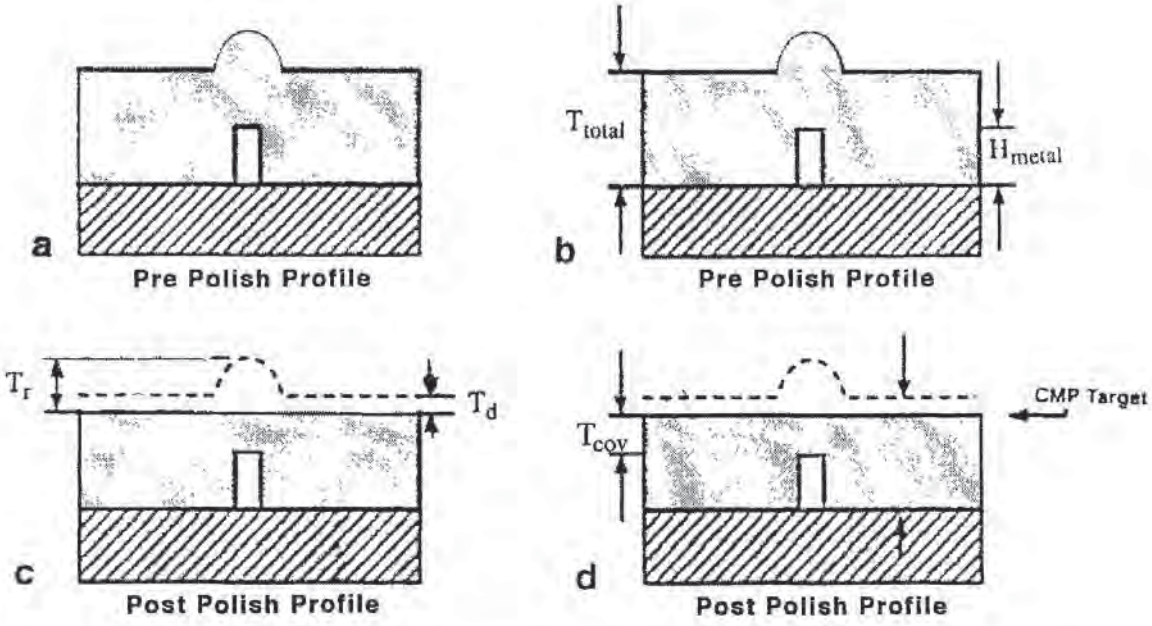


Fig. 8-17 The basic CMP process is one of reducing film-thickness selectivity over the wafer, ending with improved planarity. Here that metric is defined as *efficiency of planarization* (EOP).

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

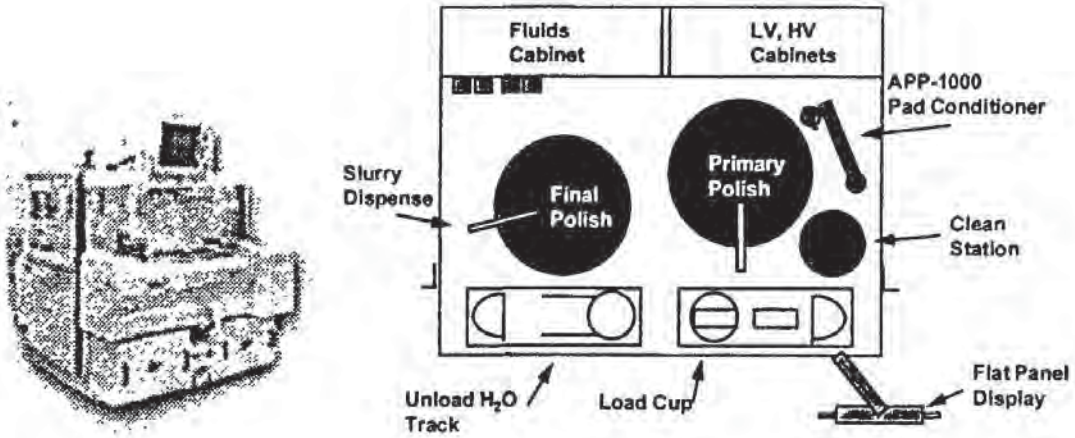


Fig. 8-18 The Speedfam 372 CMP tool.

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:

$$\mathrm{RR} = K_p P v \tag{8.13}$$

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:

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

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:

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

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:

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

and

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

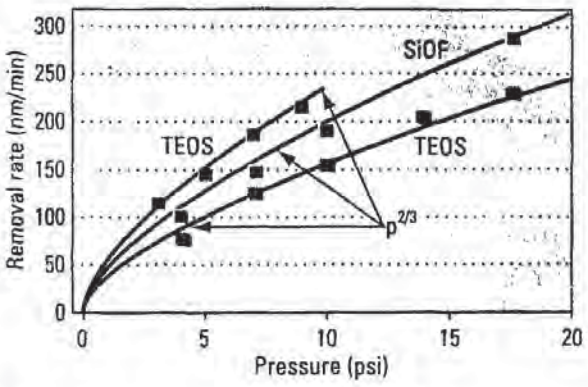


Fig. 8-19 Experimental data correlates well with the $P^{2/3}$ dependence of removal-rate on applied pressure.[16]

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.

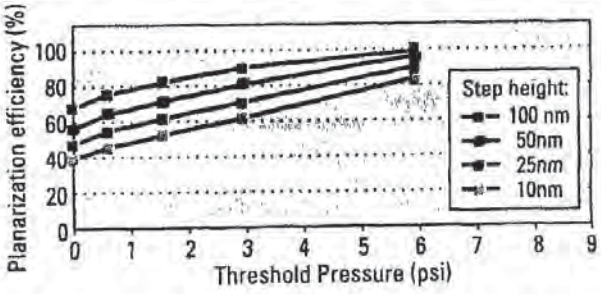


Fig. 8-20 Increasing the threshold pressure improves planarization efficiency for various final step heights (500-nm initial step height).[16]

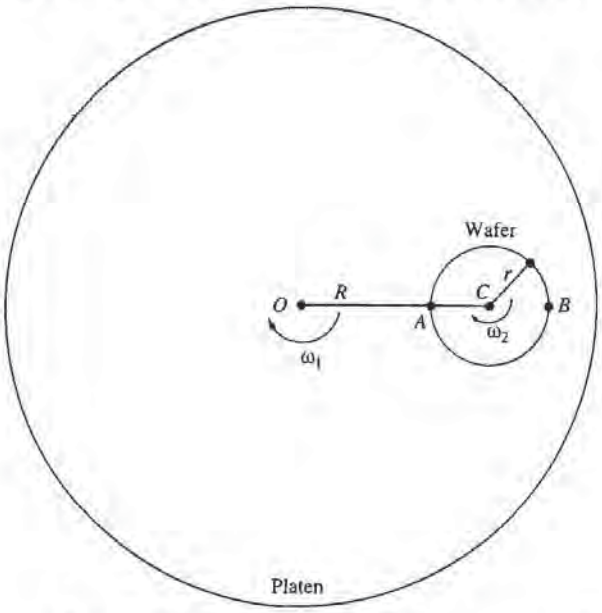


Fig. 8-21 Plan view of a single-carrier single-platen rotary-polisher, showing relative rotary motions of both the carrier and the platen.

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-

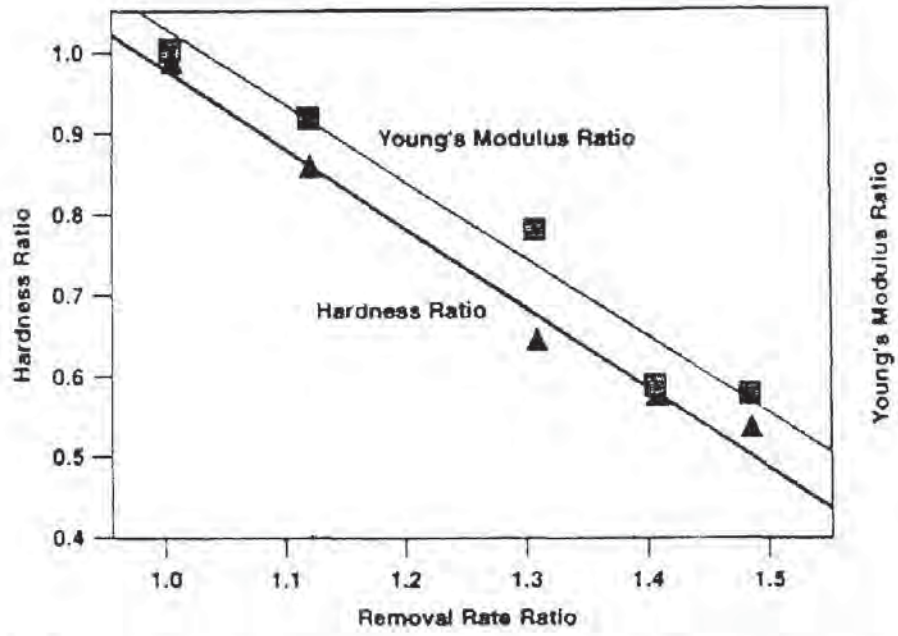
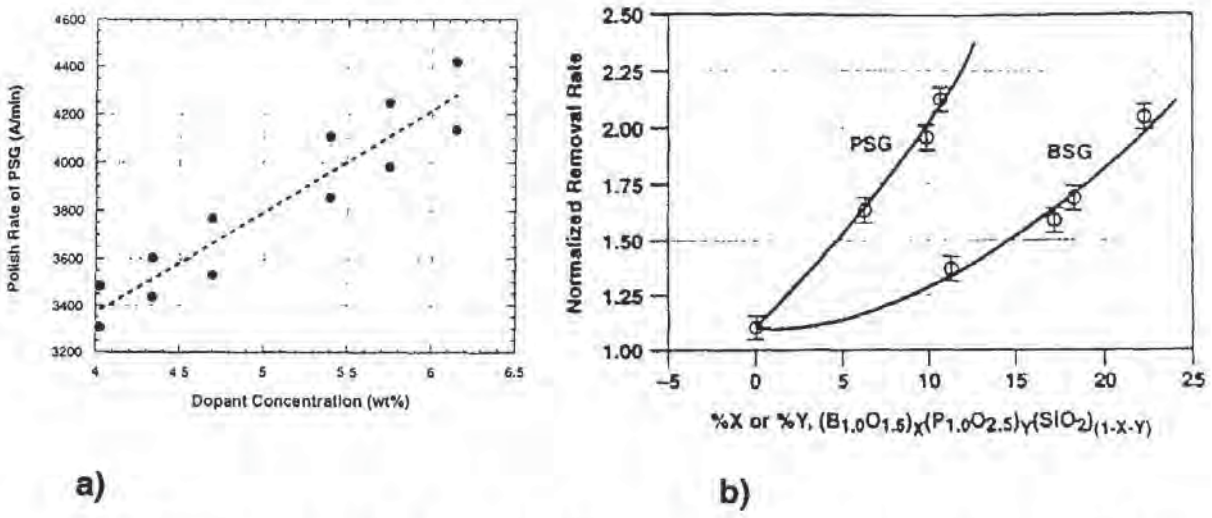


Fig. 8-22 A linear relationship exists between the CMP-removal-rate and the hardness and elastic modulus of F-SiO_2. Removal-rate, hardness, and elastic modulus are all normalized with respect to those parameters of undoped TEOS-oxides.[18]



a) b)

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

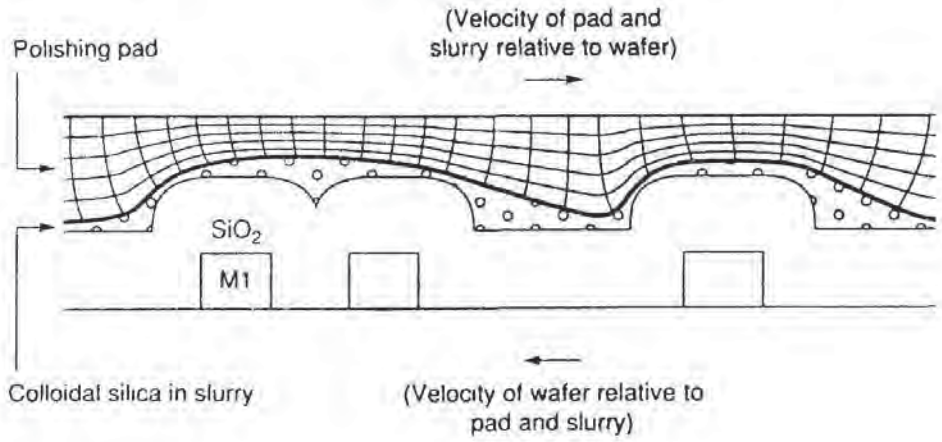


Fig. 8-24 Depiction of polishing-pad compression and localized pressures applied to topographic features. Pressures are greatest at the top-leading-edges of steps, and are lowest in the wakes of the steps.

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.

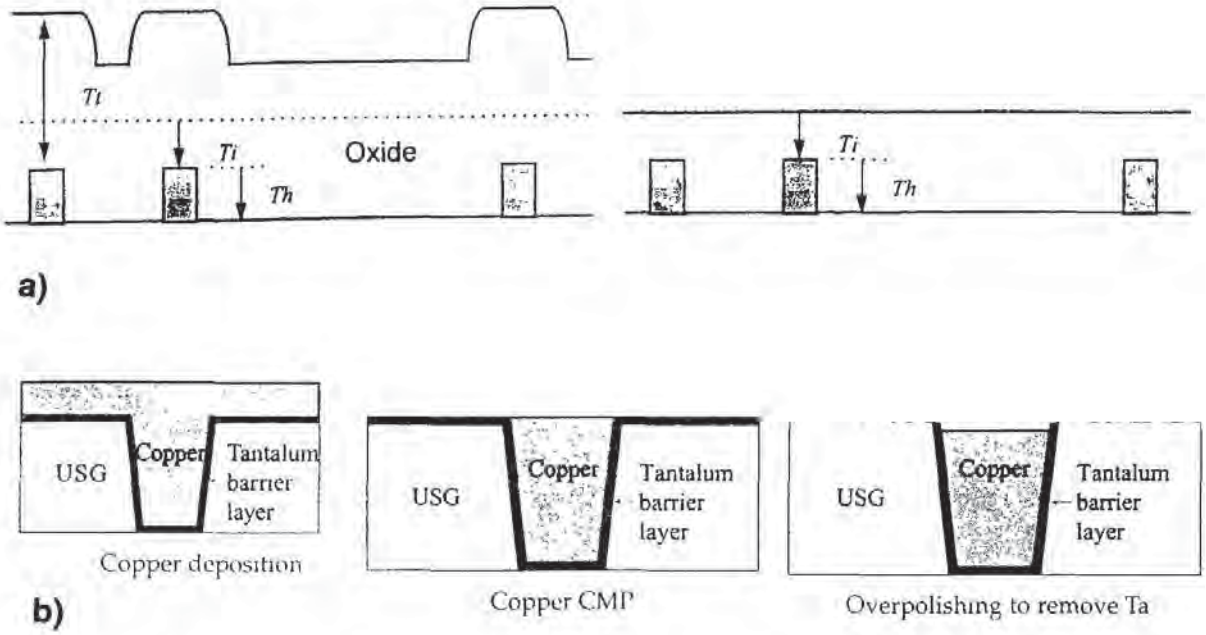


Fig. 8-25 The process of removing material by CMP falls into two categories: a) Removing homogeneous material and stopping while the removal of the surface film is incomplete (e.g., planarization of an ILD oxide); b) Removing deposited material, but continuing until all material is entirely removed on 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.

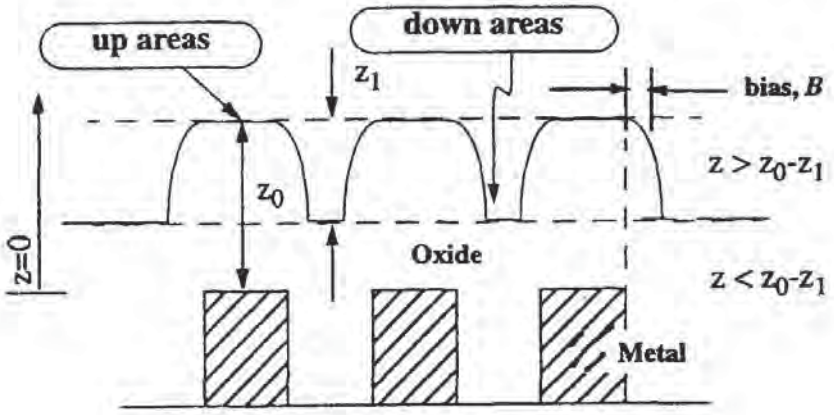


Fig. 8-26 Cross section of conformal dielectric on patterned metal, showing the "up-regions" and "down-regions" of the dielectric film.

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.

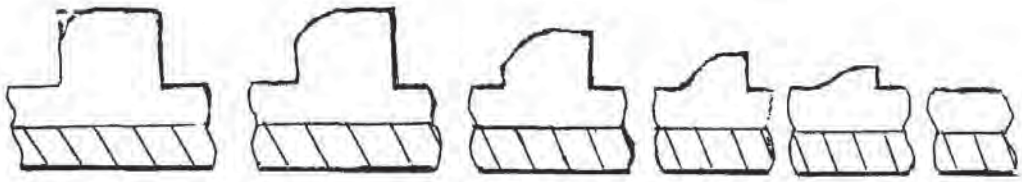


Fig. 8-27 A depiction of how the leading edge of a feature of an oxide step (i.e., the edge that is pressed against the pad by the combined movements of the carrier and platen), is eroded by CMP.

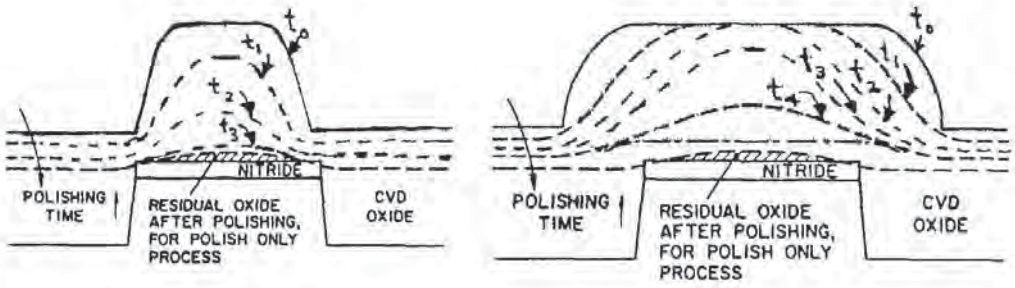


Fig. 8-28 Schematic cross-section of a narrow and wide feature of oxide being polished by CMP, as a function of time. Since the leading edges are polished most rapidly (and the feature continually rotates, so that all edges become the leading edge for the same amount of time), the narrow feature is planarized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

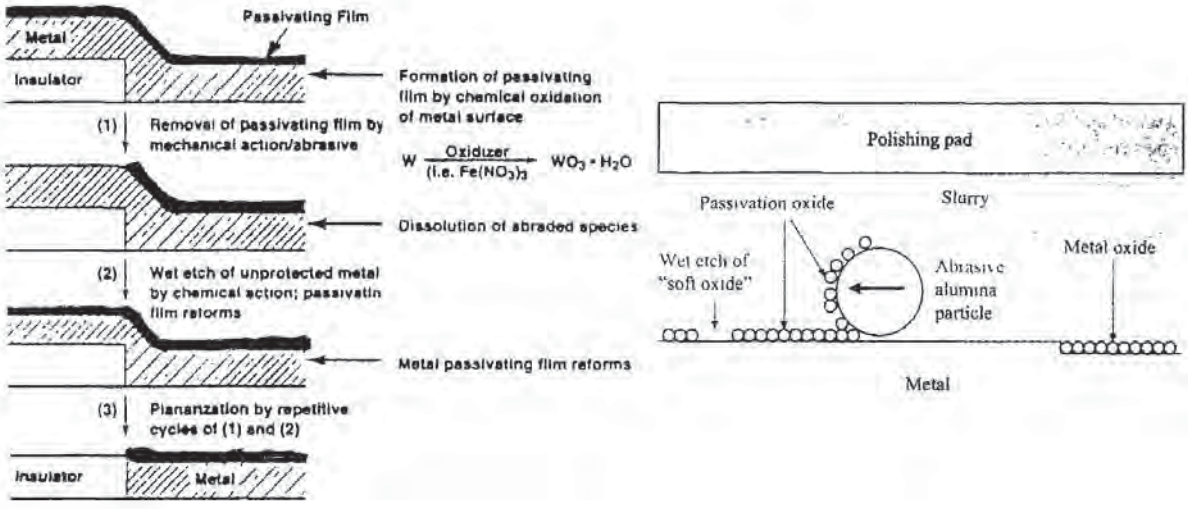


Fig. 8-29 Schematic representation of the model of CMP of metal films.[26]

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-

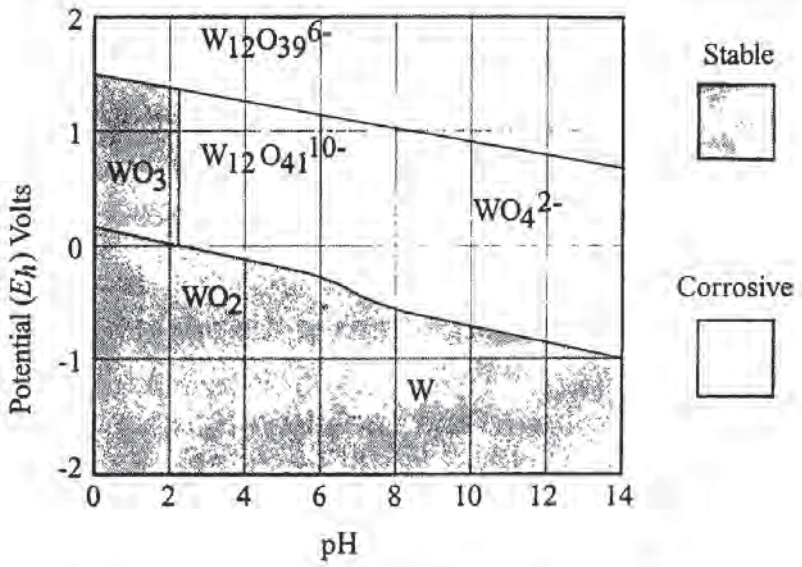


Fig. 8-30 Tungsten solubility versus solution-pH (Pourbaix) diagram.[27]

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 overpolish 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.

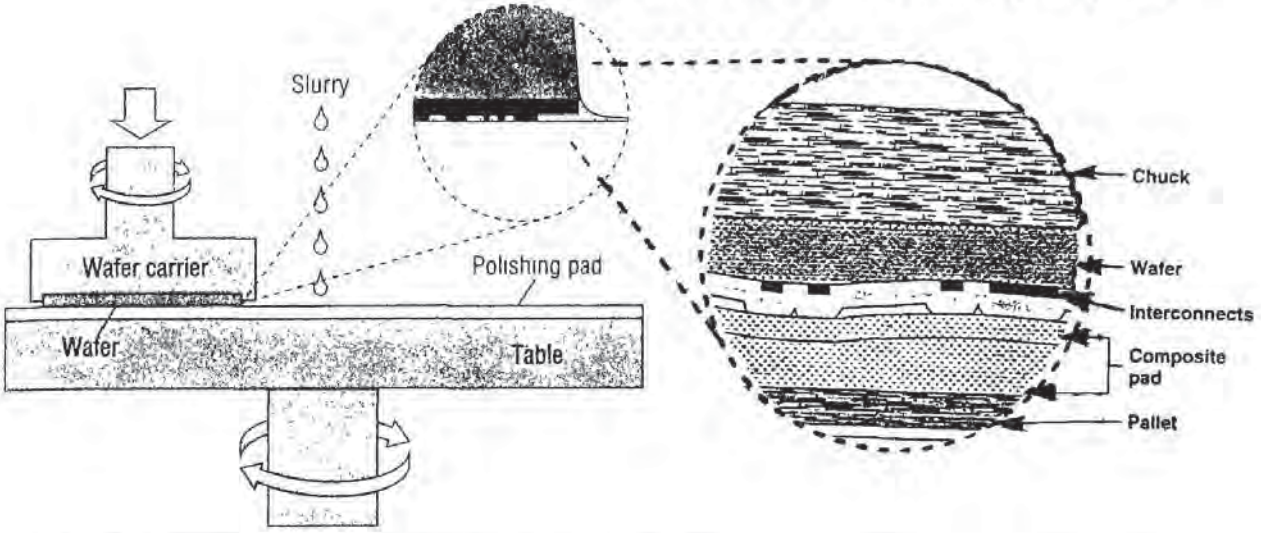


Fig. 8-31 Schematic diagram of a wafer being polished by CMP.

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.

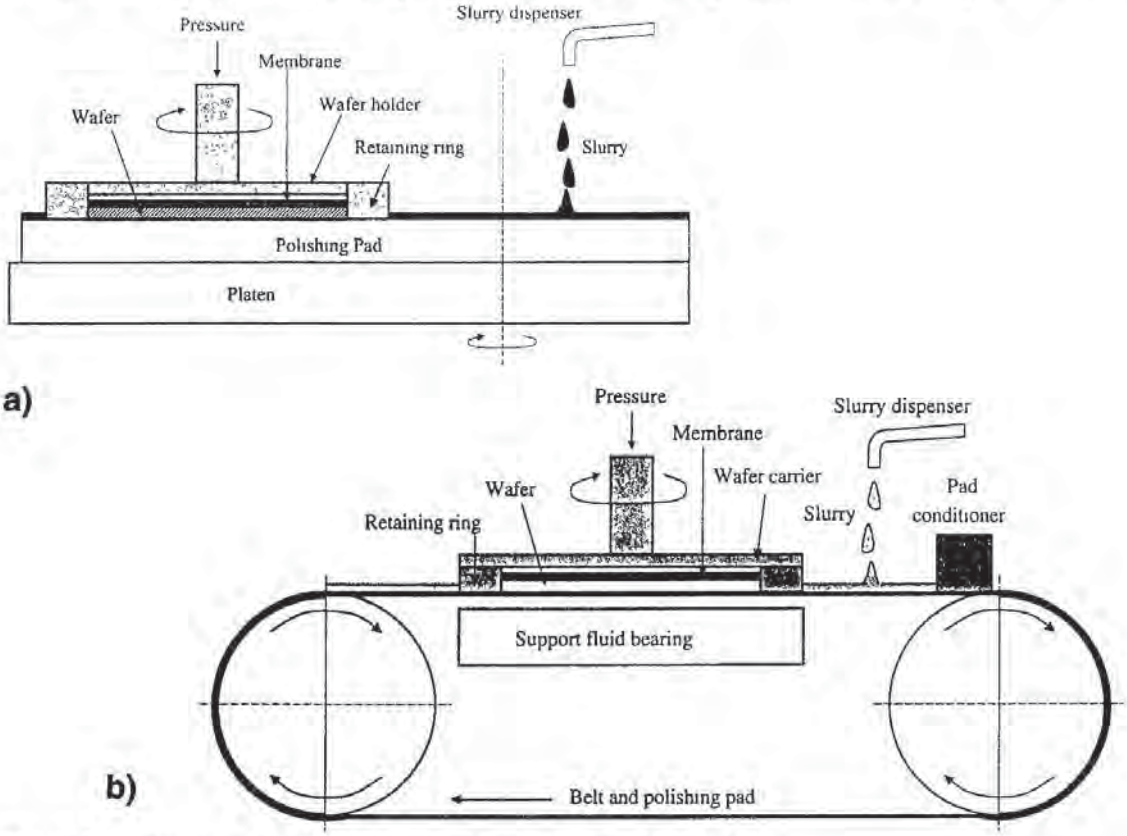


Fig. 8-32 CMP tool configurations: a) Rotary tool; b) Linear tool.

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

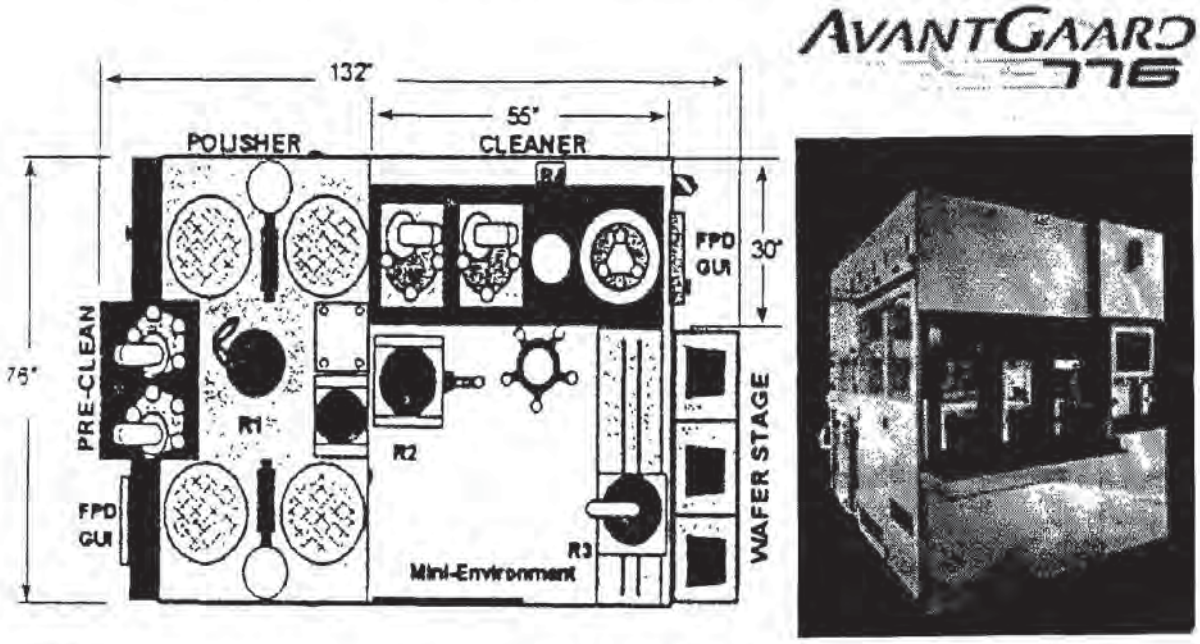


Fig. 8-33 Plan view of an IPEC *Avanti* 776® CMP tool. Courtesy of Speedfam/IPEC.

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,

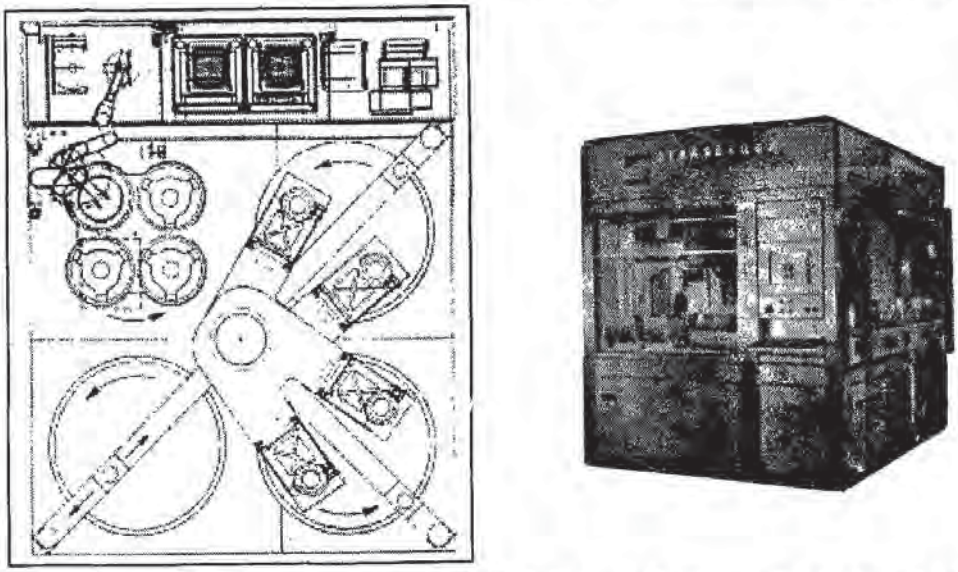


Fig. 8-34 Schematic drawing of the Symphony® CMP tool. Courtesy of Strasbaugh Corp.

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

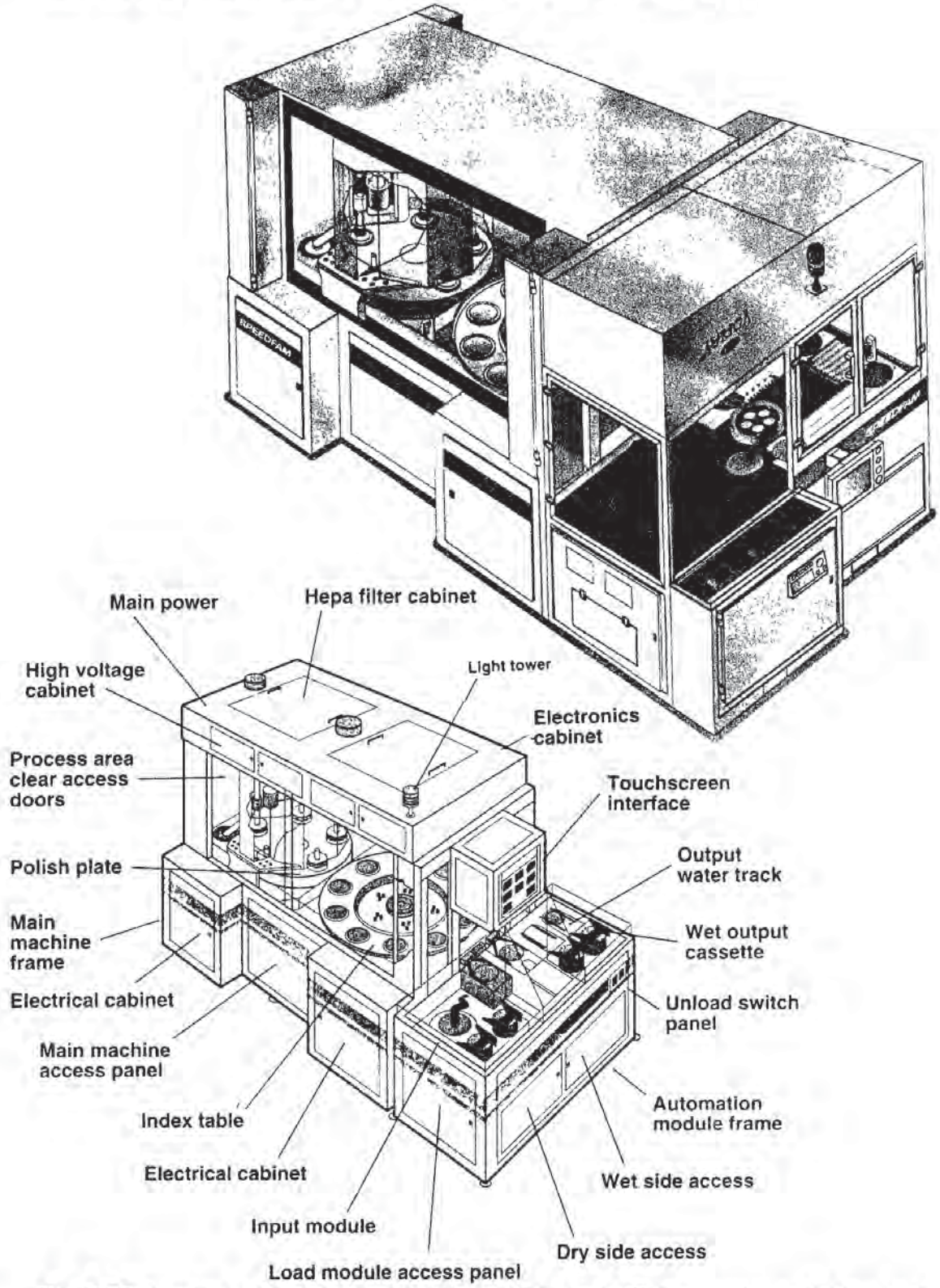


Fig. 8-35 Detailed view of the SpeedFam-IPEC Auriga® five-carrier polishing tool. Courtesy of SpeedFam-IPEC, Inc.

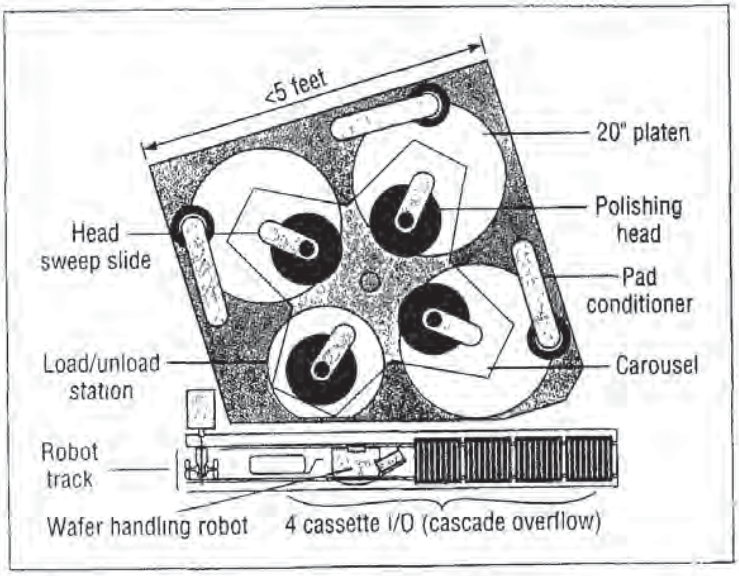


Fig. 8-36 Architecture of the Mirra® three-platen CMP tool. Courtesy of Applied Materials, Inc.

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

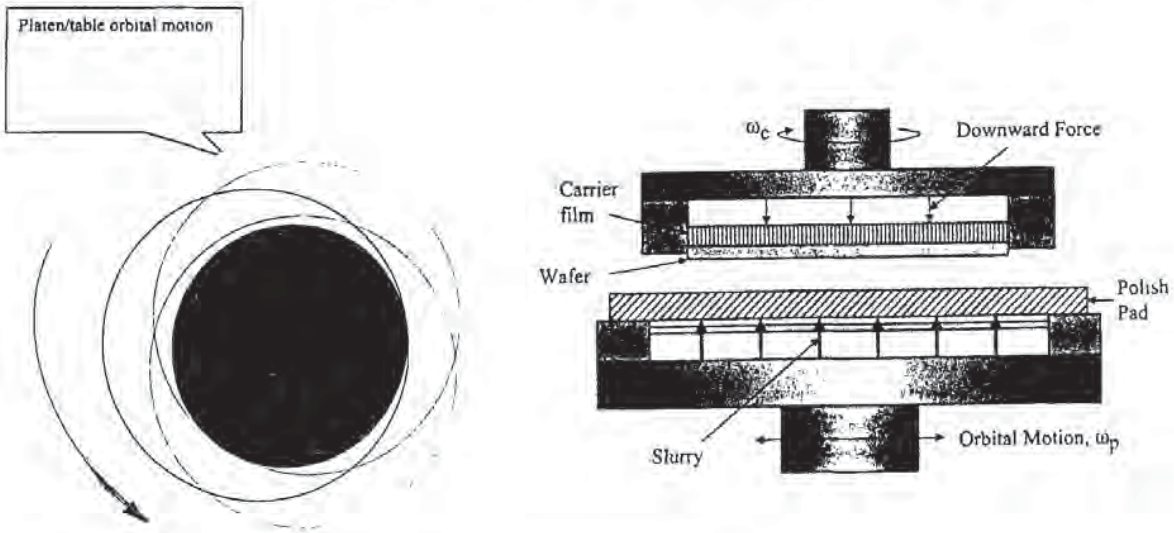


Fig. 8-37 Schematic drawing of orbital platen motion polisher.

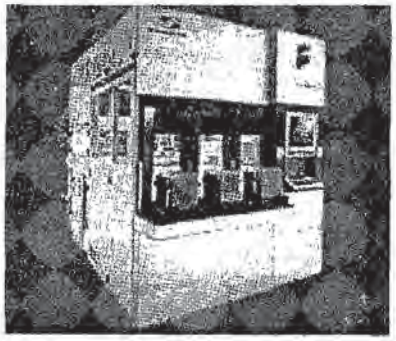


Fig. 8-38 Photograph of the Momentum,[10] an orbital platen motion polisher. Courtesy of SpeedFam-IPEC, Inc.

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

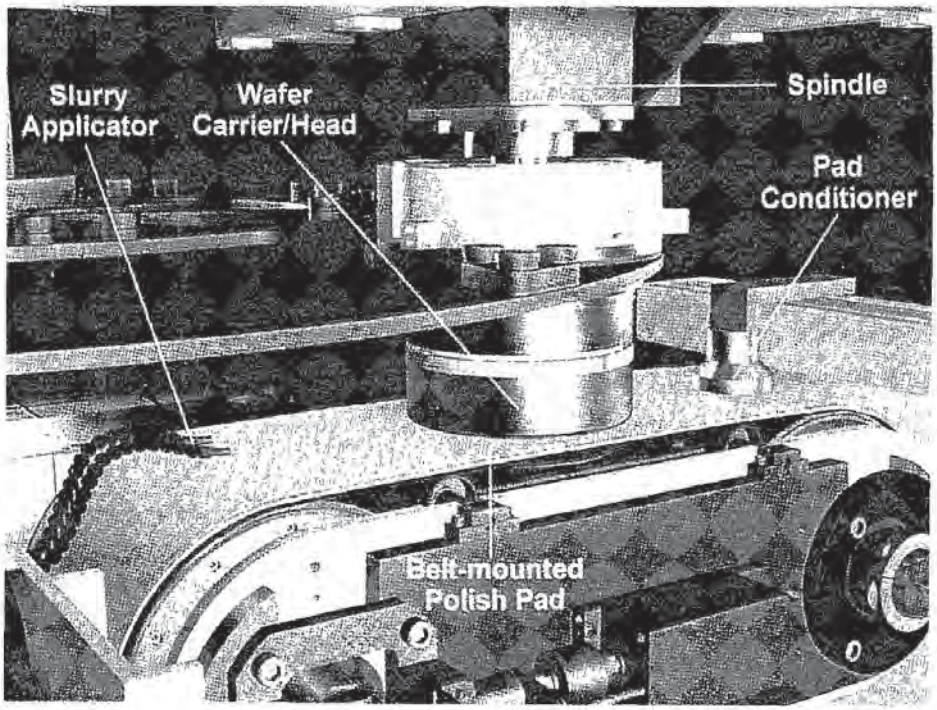


Fig. 8-39 Schematic drawing of the Teres® linear CMP tool. Courtesy Lam Research, Inc.

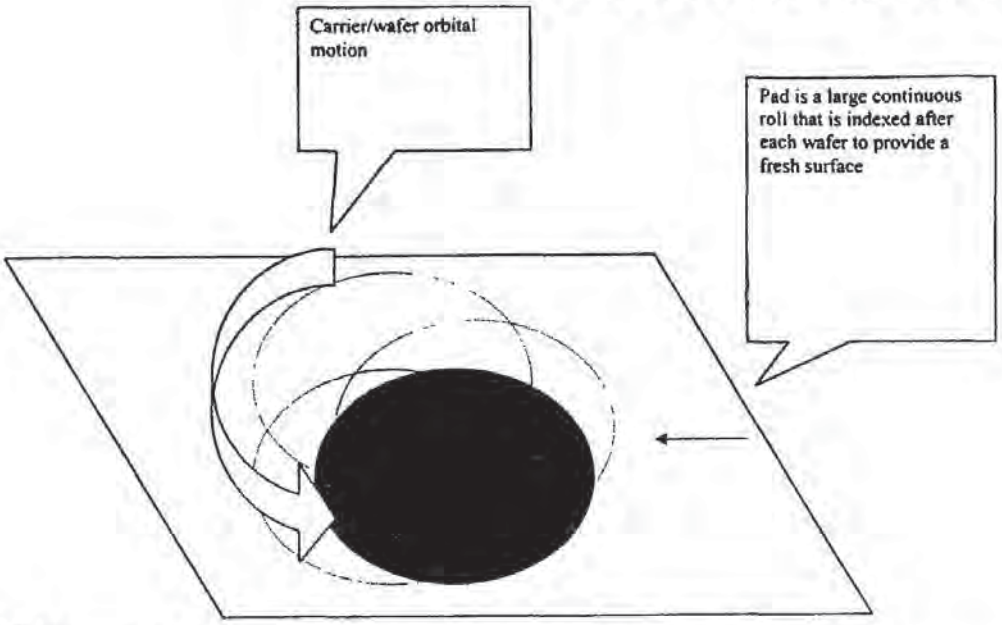


Fig. 8-40 Schematic drawing of orbital carrier-motion with indexing pad.

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.

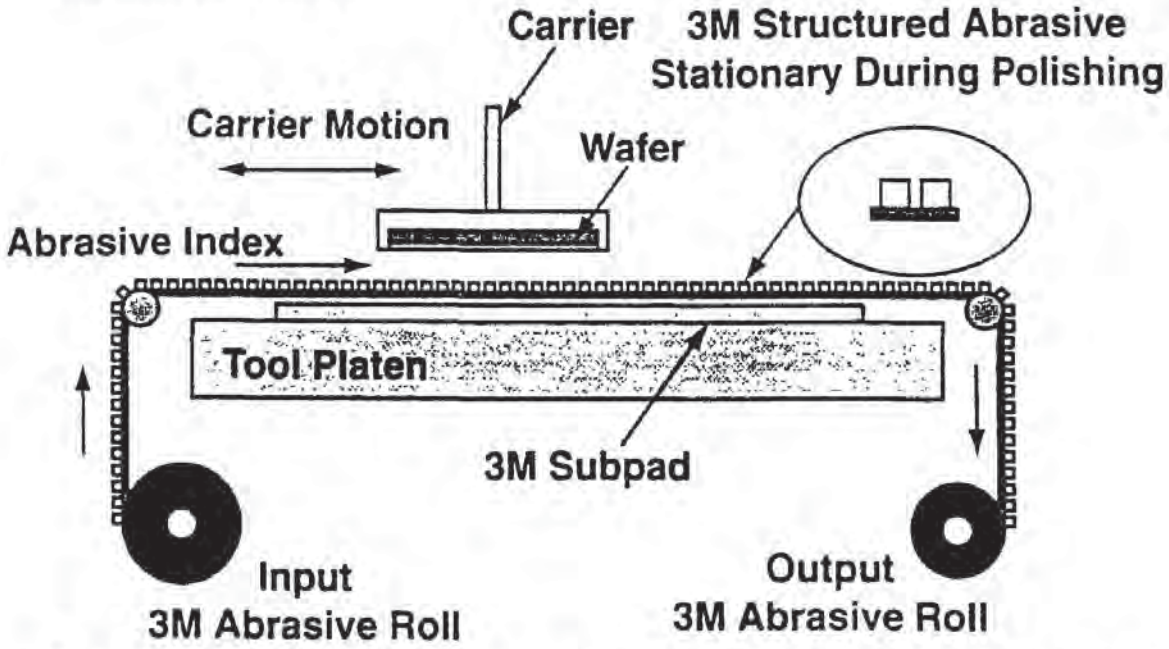


Fig. 8-41 Fixed-table with roll-to-roll pad. Carrier rotates and also executes linear motion.

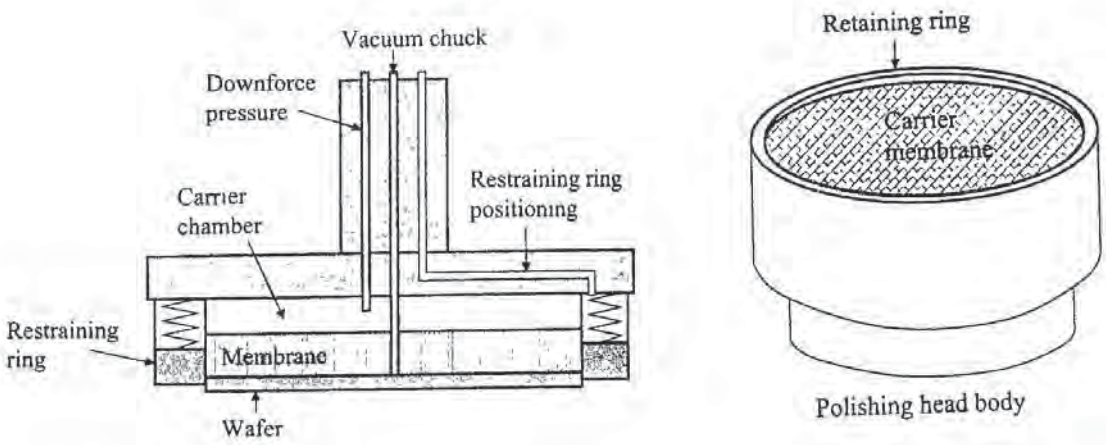


Fig. 8-42 Schematic drawing of the components of a CMP wafer-carrier.

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

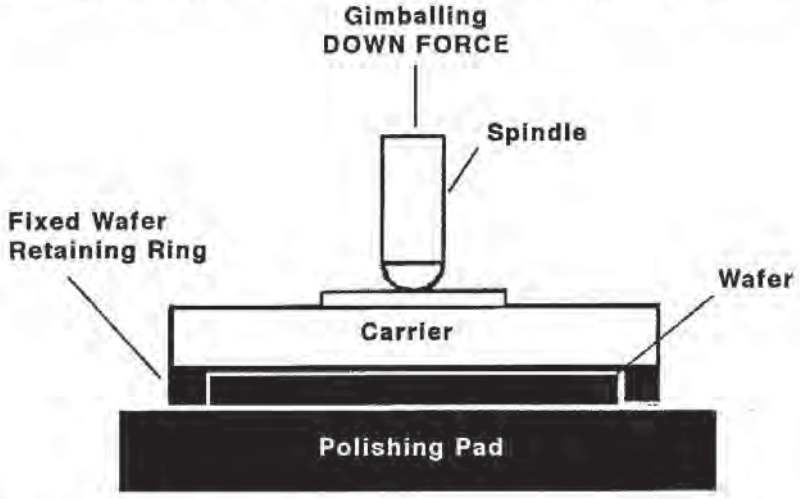


Fig. 8-43 Carrier being subjected to gimballing down force.

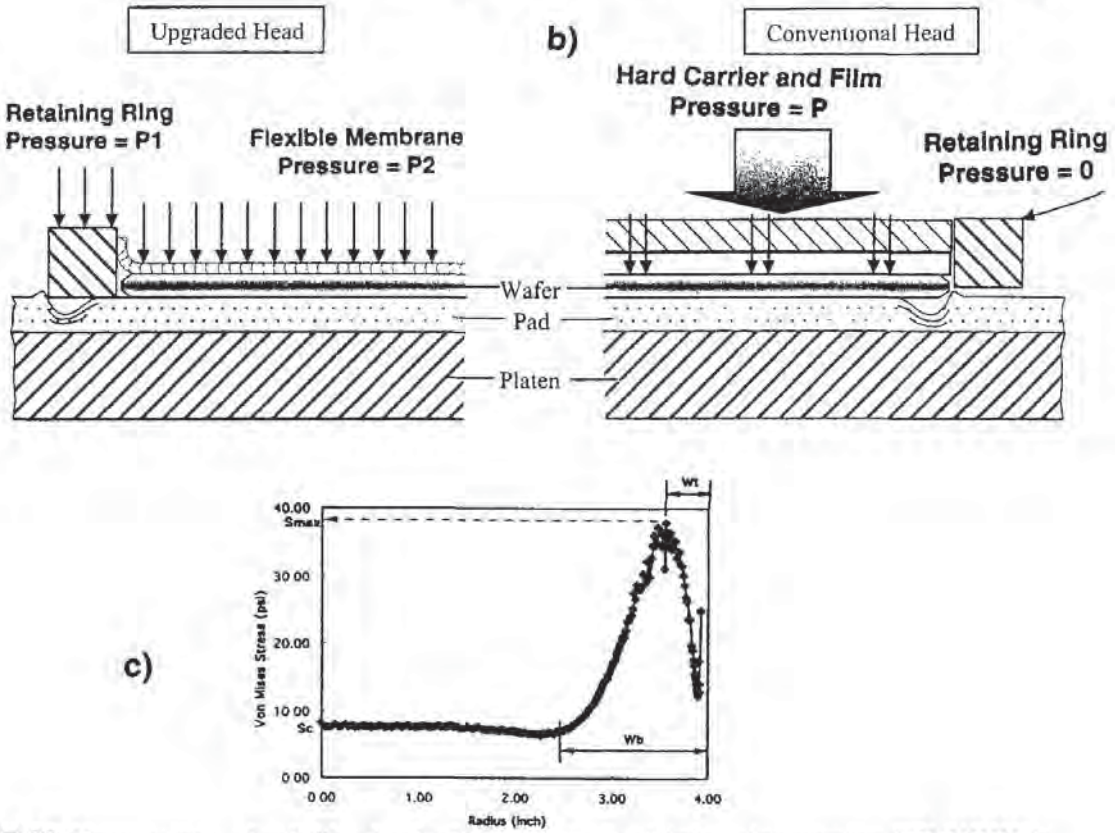


Fig. 8-44 Comparison of: a) film-backed/non-contact retaining-ring carrier (FB-NRRC); and b) bladder-backed/contact retaining-ring carrier (BB-CRRC); c) distribution of pressure profile at wafer edge for FB-NRRC.

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

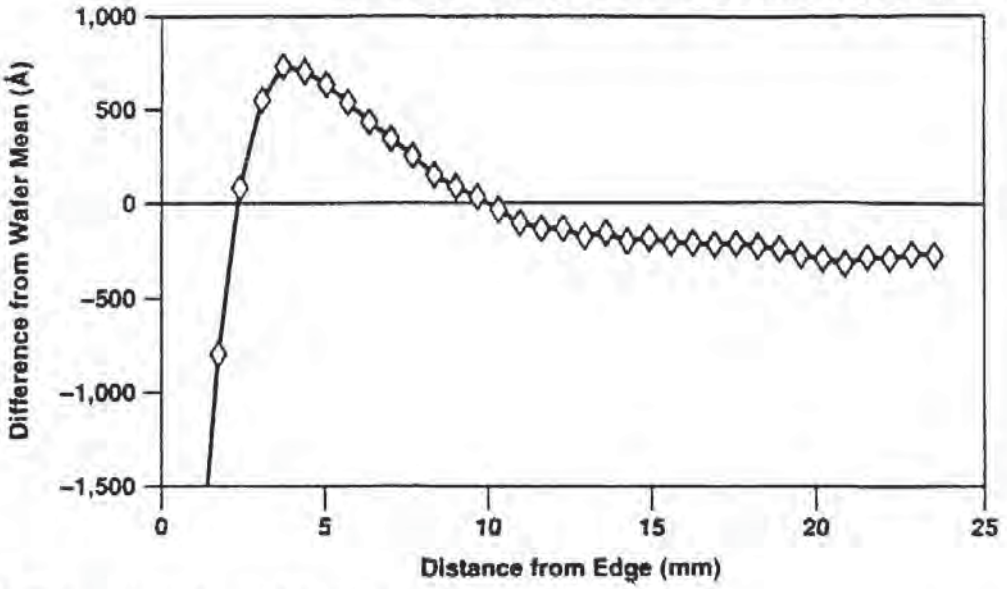


Fig. 8-45 Edge effect observed in CMP.

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

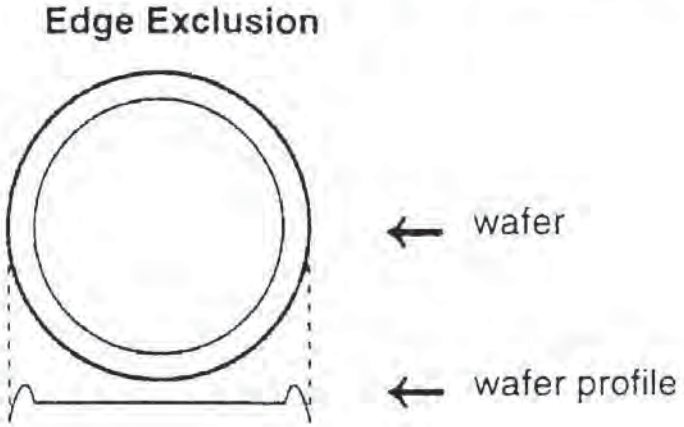


Fig. 8-46 Exclusion ring.

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

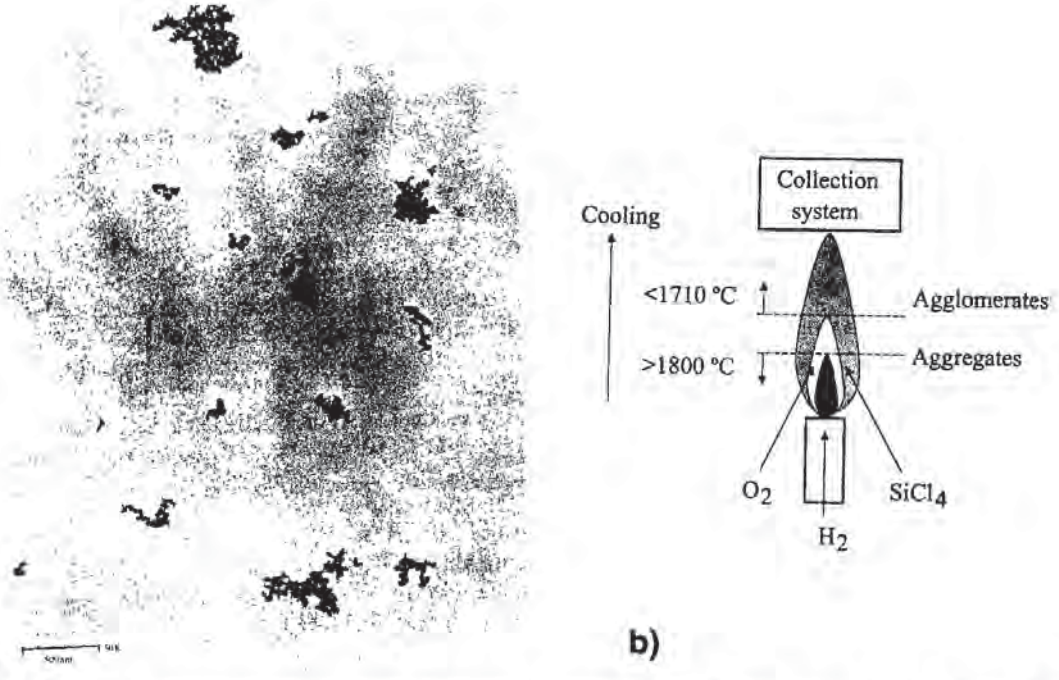


Fig. 8-47 a) SEM of fumed-silica slurry-particles. Courtesy Rodel; b) Fumed-silica formation.

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

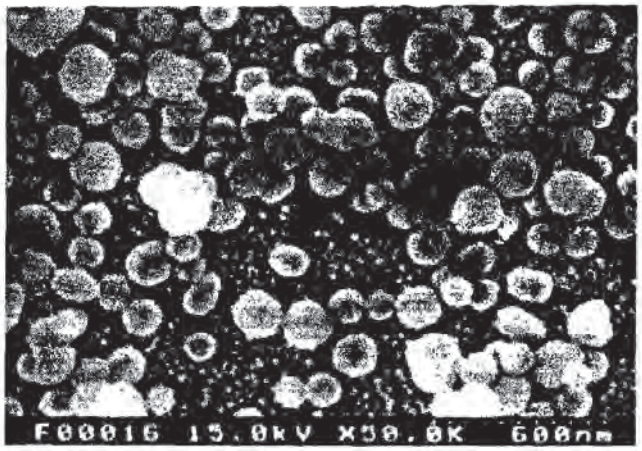


Fig. 8-48 SEM of colloidal-silica slurry particles. Courtesy of Fujimi Corporation

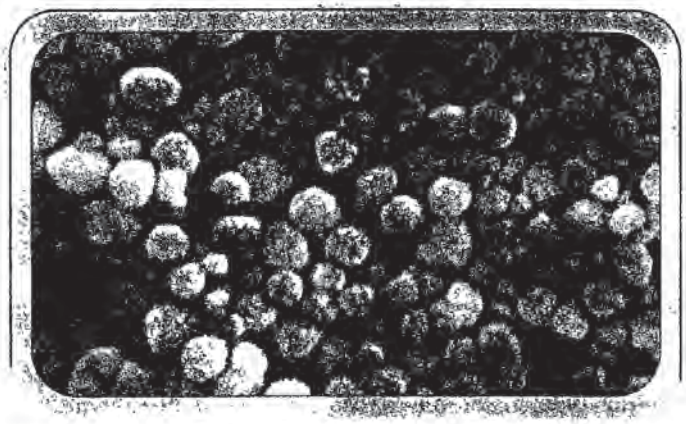


Fig. 8-49 SEM of cerium oxide slurry particles. Courtesy of Fujimi Corporation.

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-

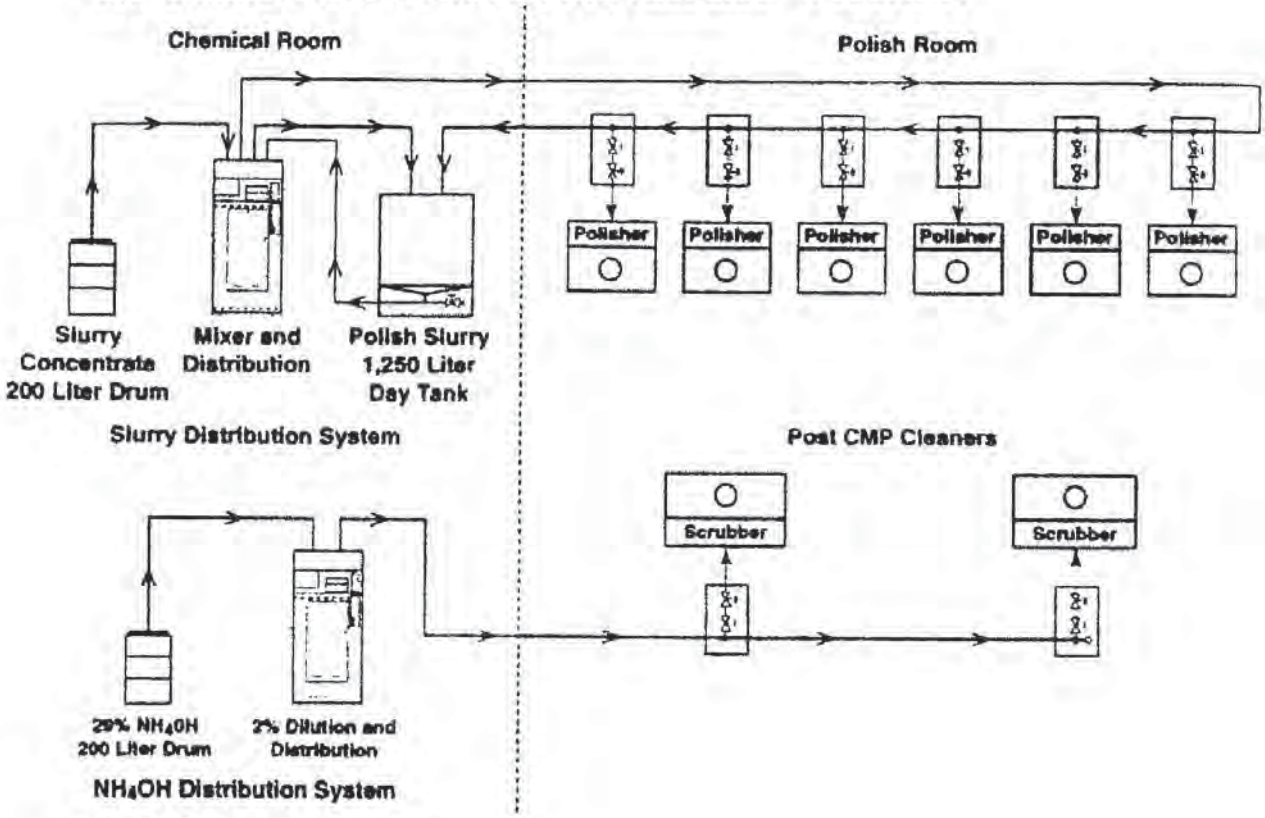


Fig. 8-50 Slurry distribution system.

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

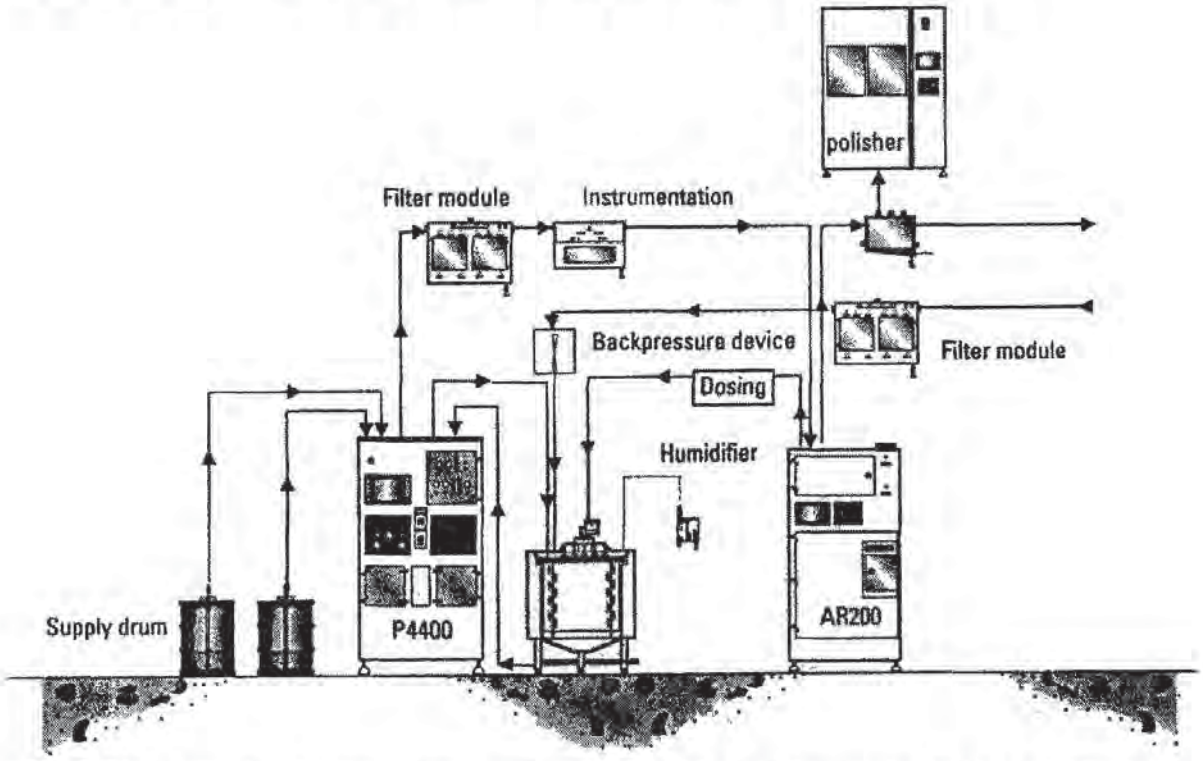


Fig. 8-51 Schematic of a slurry management-system with metrology, auto-dosing and filtration.

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

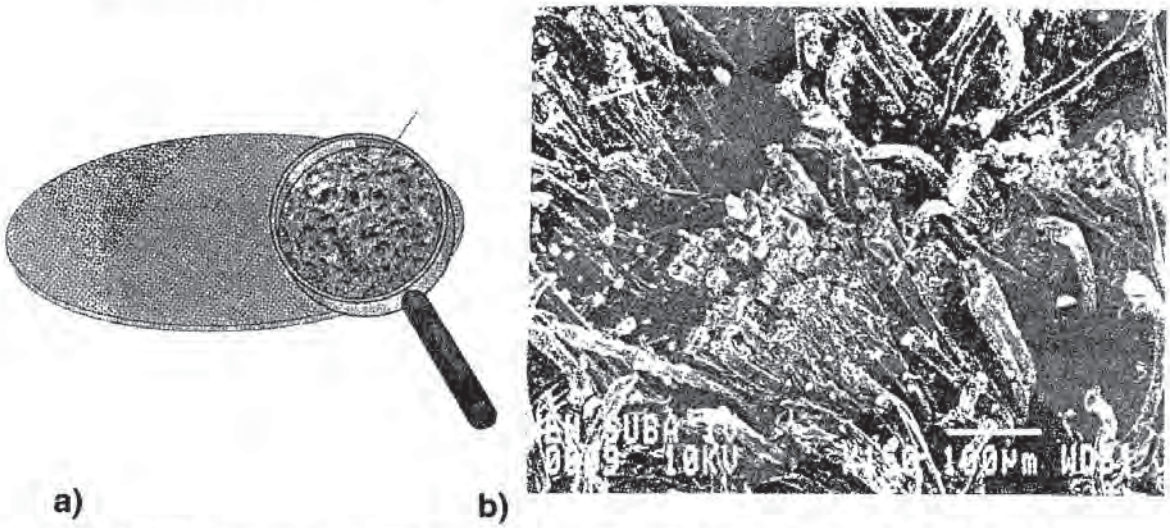


Fig. 8-52 a) Schematic of polishing pad; b) SEM of a new Rodel Suba IV® Pad. Courtesy of Rodel Inc.

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

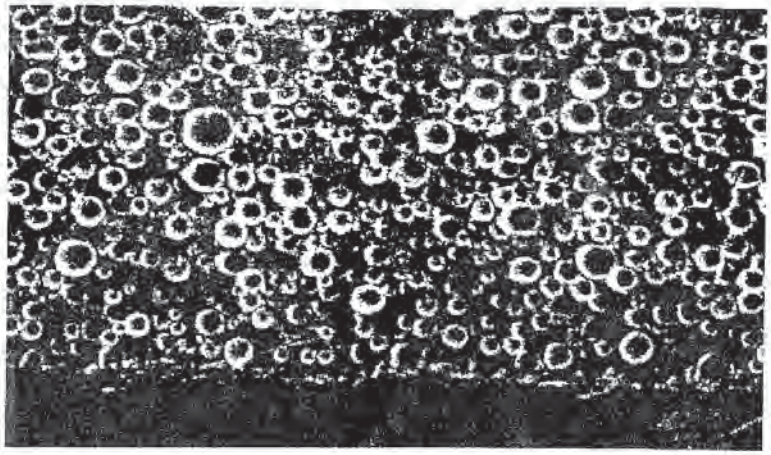


Fig. 8-52 SEM of a new Rodel 1000 IC Pad. Courtesy of Rodel Inc.

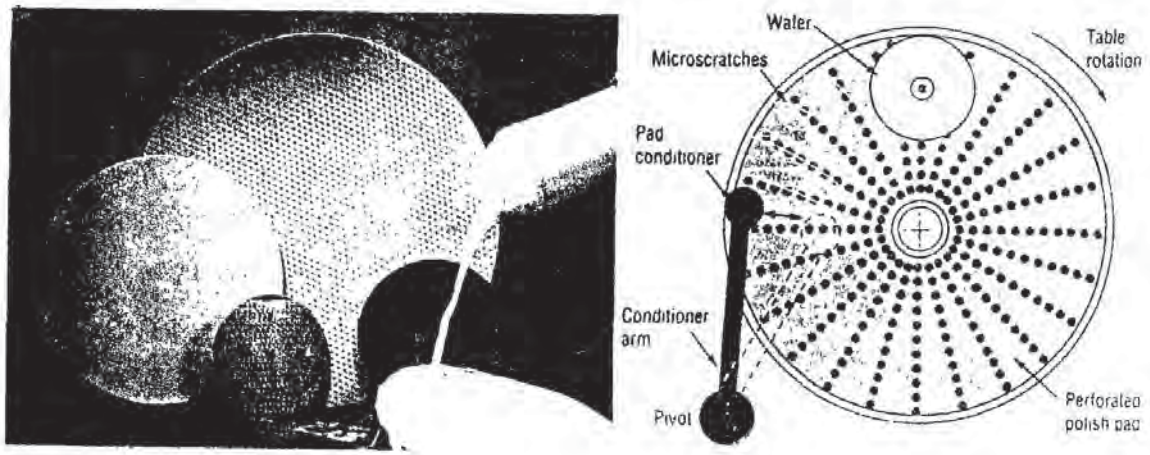


Fig. 8-54 Photograph of polishing pads, showing pad perforations. Courtesy of Rodel Inc.

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

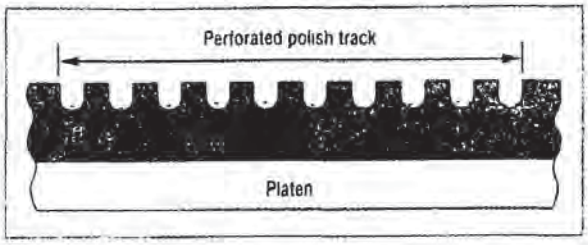


Fig. 8-55 Polishing pad cross-section, showing grooves in pad surface.

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.

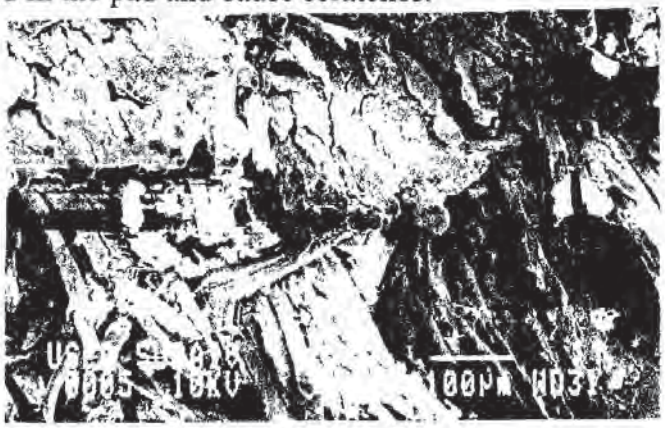


Fig. 8-56 Surface of a glazed pad (Rodel's Suba IV® pad) after polishing.

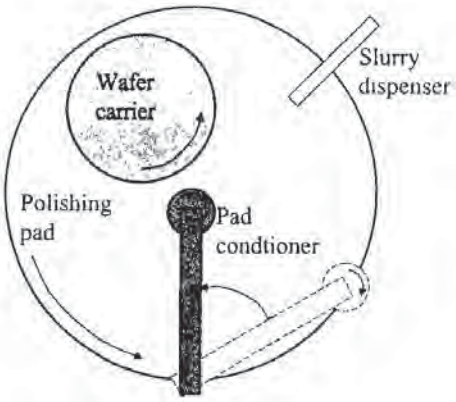


Fig. 8-57 Schematic drawing of pad conditioning. Reprinted with permission of Solid State Technology. Published by PennWell.

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

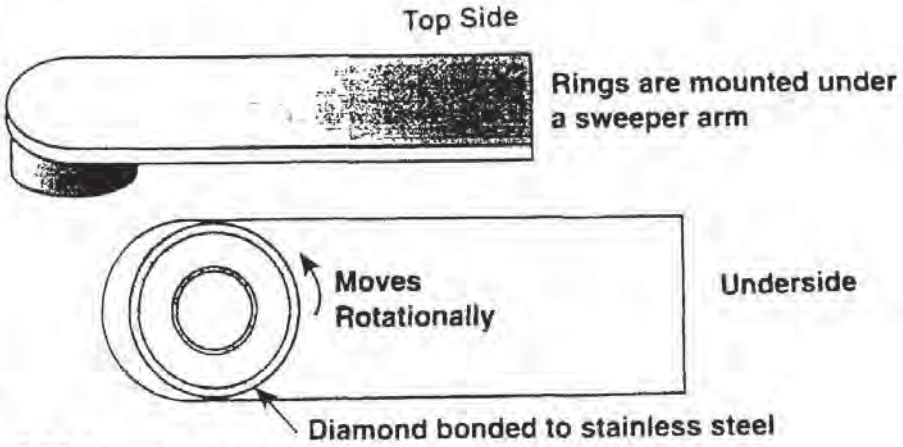


Fig. 8-58 Diamond embedded ring pad conditioner.

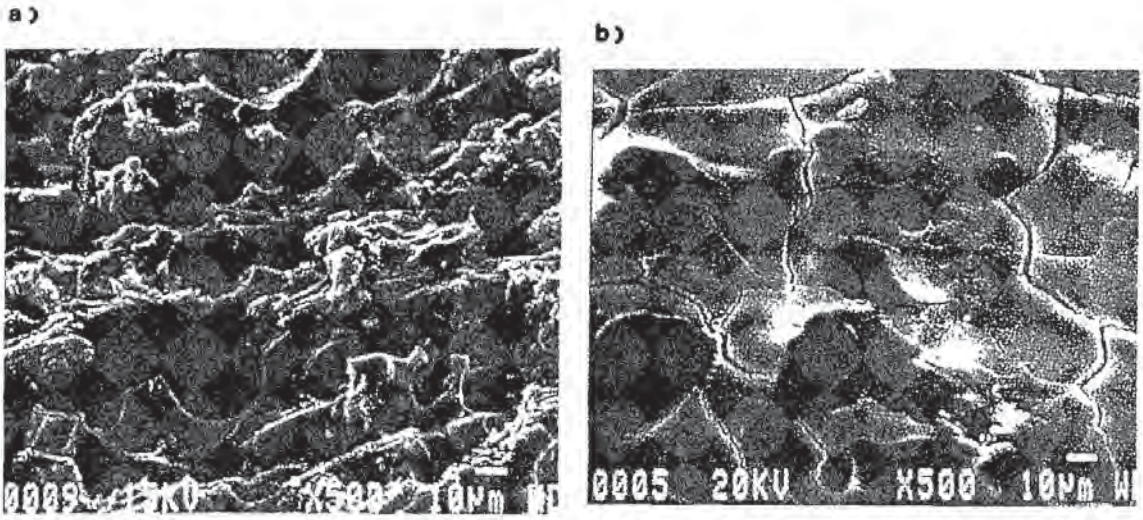


Fig. 8-59 (a) Polishing pad cross section. (b) Effect of pad conditioning.[47] Reprinted with permission of Solid State Technology. Published by PennWell.

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

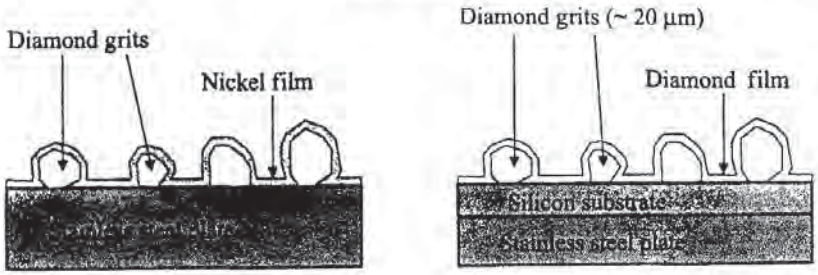


Fig. 8-60 Surface of pad conditioner showing bonded diamond grit.

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.

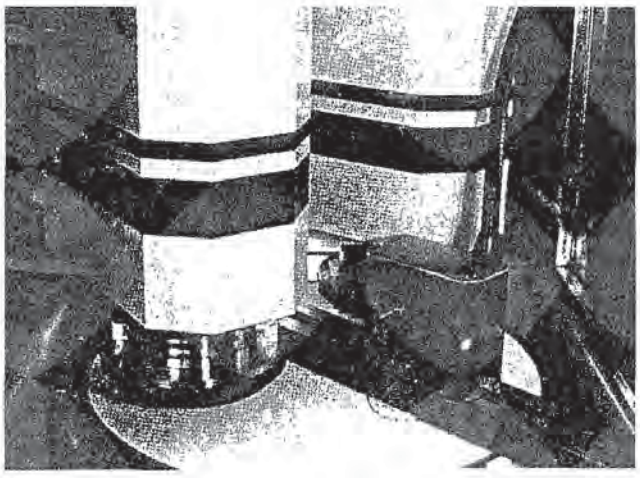


Fig. 8-61 In situ pad conditioner. Courtesy of Strasbaugh Inc.

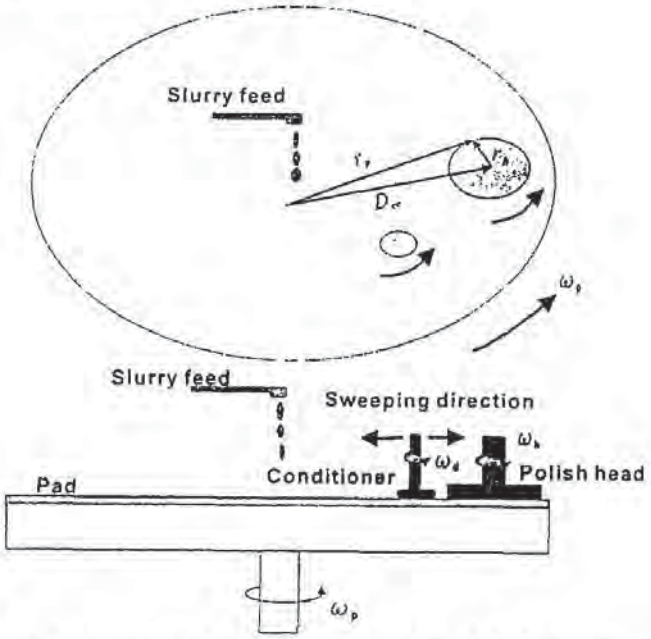


Fig. 8-62 *In situ* pad conditioner.

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.

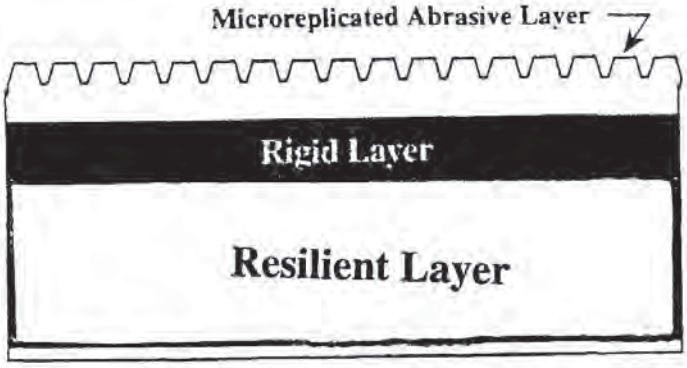


Fig. 8-63 Slurry-free (Fixed abrasive) pad structure.

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

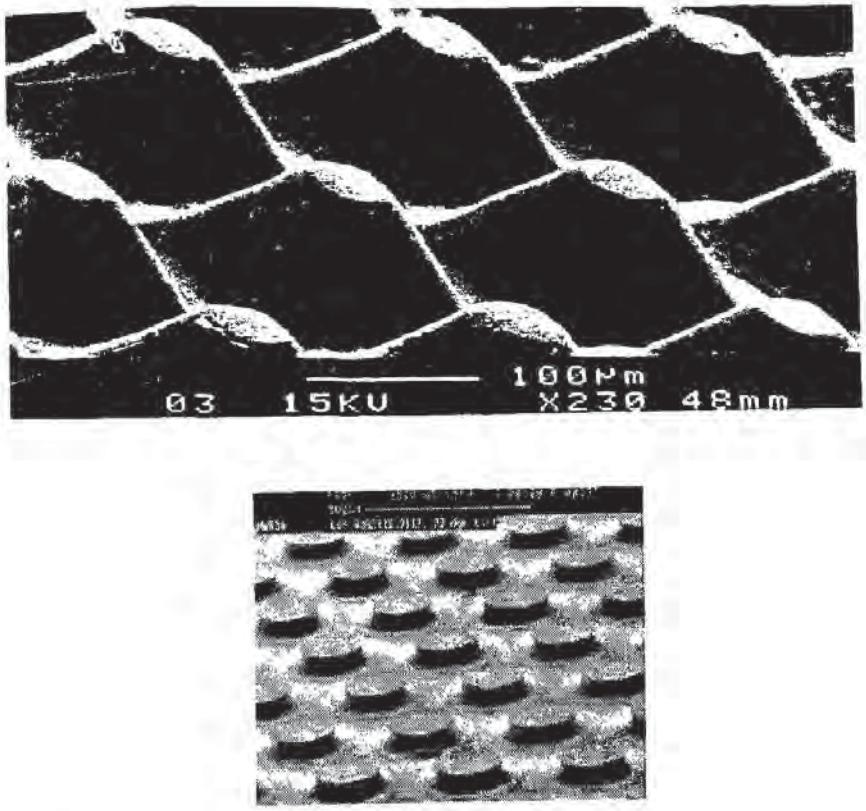


Fig. 8-64 Examples of slurry-free pad surface structures. Courtesy of 3M Co.

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.

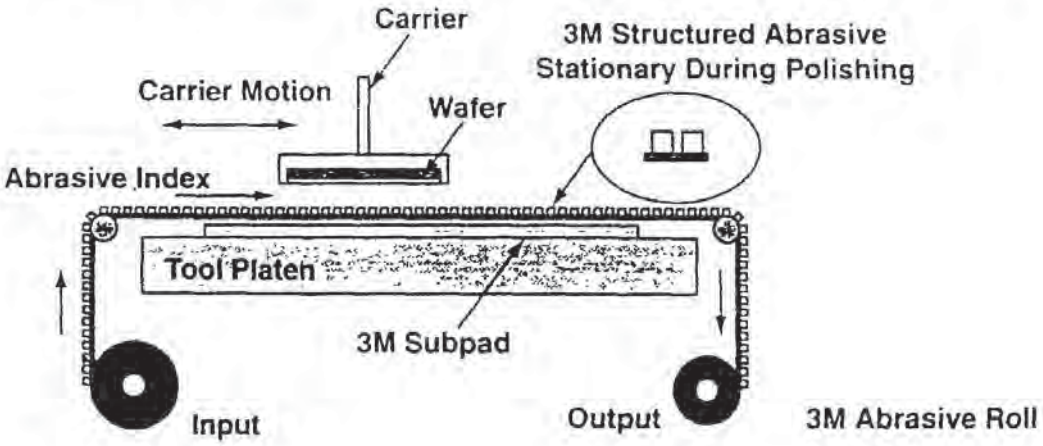


Fig. 8-65 Schematic of linear web format CMP tool, with indexing, slurry-free pad.

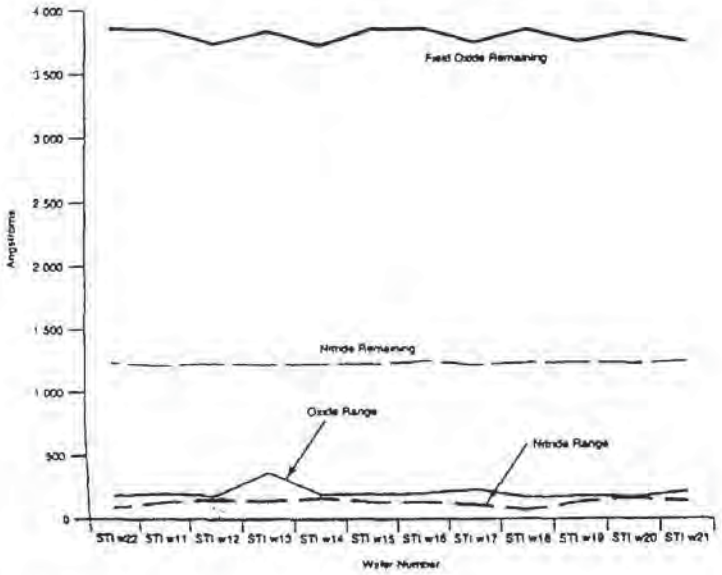


Fig. 8-66 Fixed abrasive polishing results on Linear CMP tool (roll-to-roll pad). Test performed on the Obsidian 501 CMP tool at the 3M Applications Lab. Courtesy 3M Co.

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

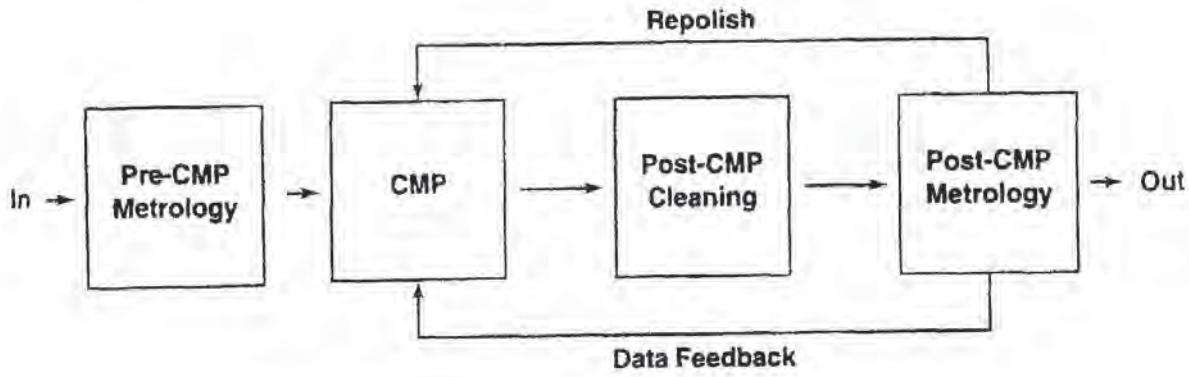


Fig. 8-67 *In-line* endpoint detection method used in CMP (i.e., polishing, followed by measurement to confirm thickness - and re-polishing if necessary - is the cycle used with *in-line* endpoint detection).

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

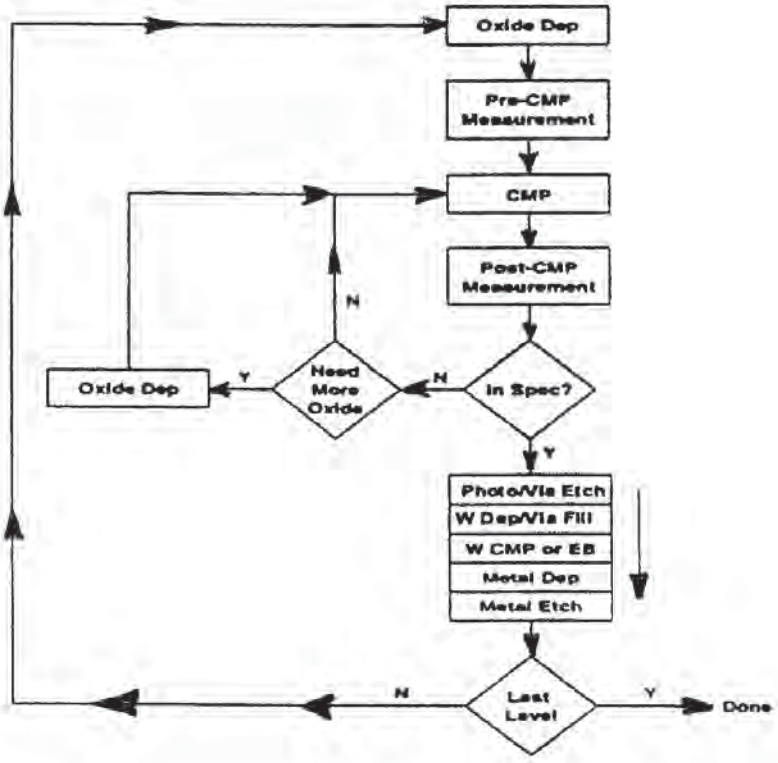


Fig. 8-68 Implementing the *in-line* endpoint-detection method used in CMP.

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.

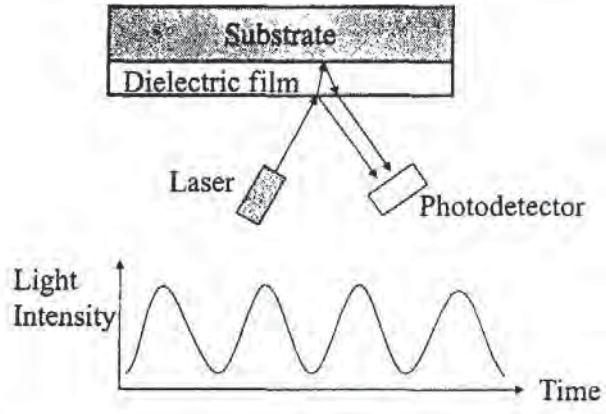


Fig. 8-69 *Wafer-frontside laser-interferometry* endpoint-detection for dielectric CMP.

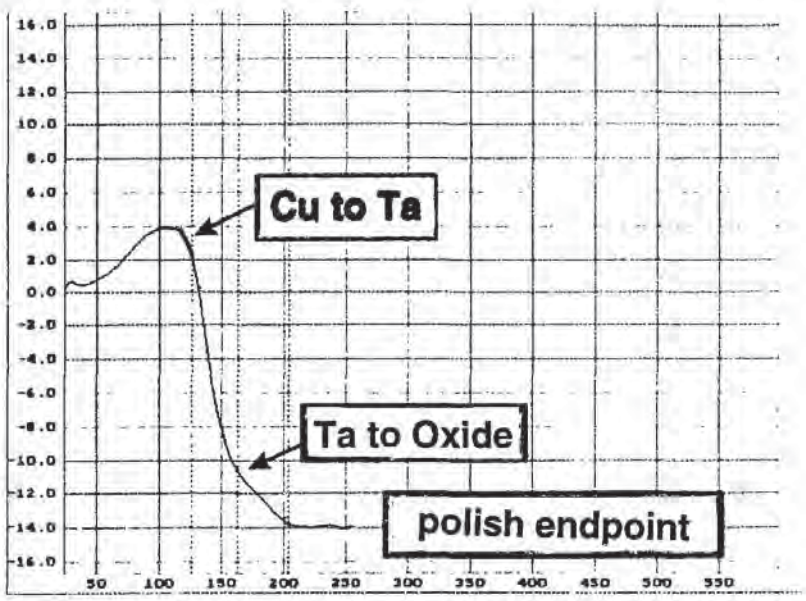


Fig. 8-70 Typical reflectance curve for polishing a Cu/Ta/oxide film.

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

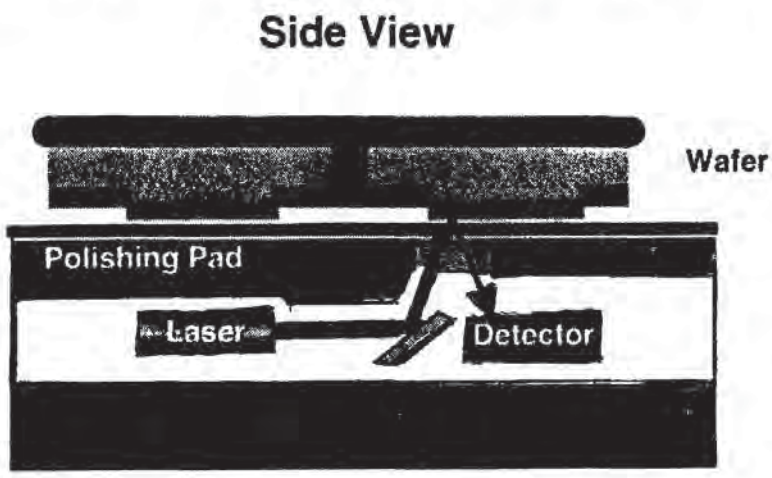


Fig. 8-71 Schematic of the Mirra® laser-reflectometer light path. Courtesy Applied Materials.

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.

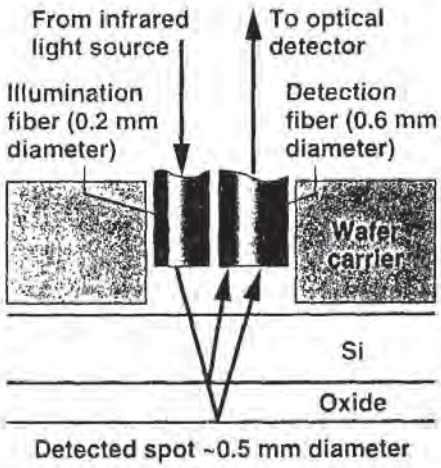


Fig. 8-72 Schematic of an optical sensor embedded in a wafer carrier.[57] Wafer is backlighted with IR light to measure films on the front surface. Reprinted with permission of Semiconductor International.

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®).

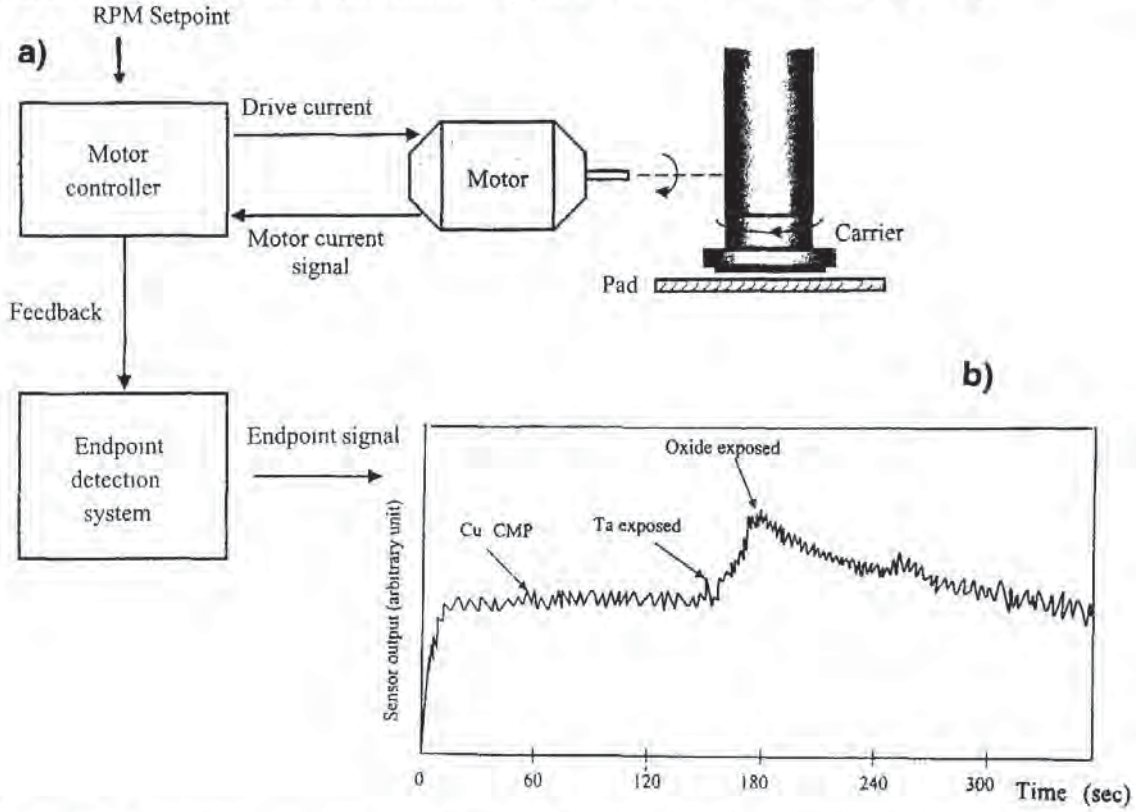


Fig. 8-73 a) Motor current endpoint detection. b) Motor current output during Cu/Ta/oxide CMP process. Courtesy of Aplex.

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

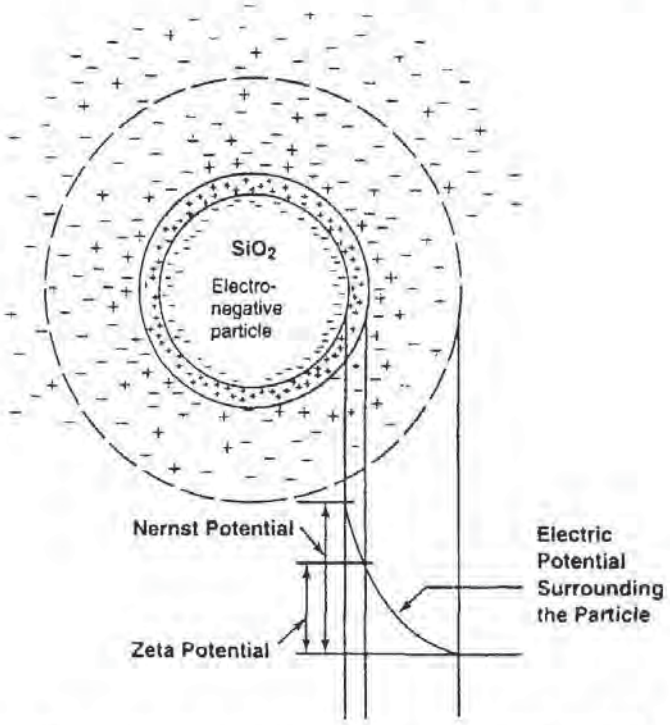


Fig. 8-74 Slurry chemistry electrostatic layers surrounding an SiO_2 particle (including a depiction of the electric dipole layer).

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.

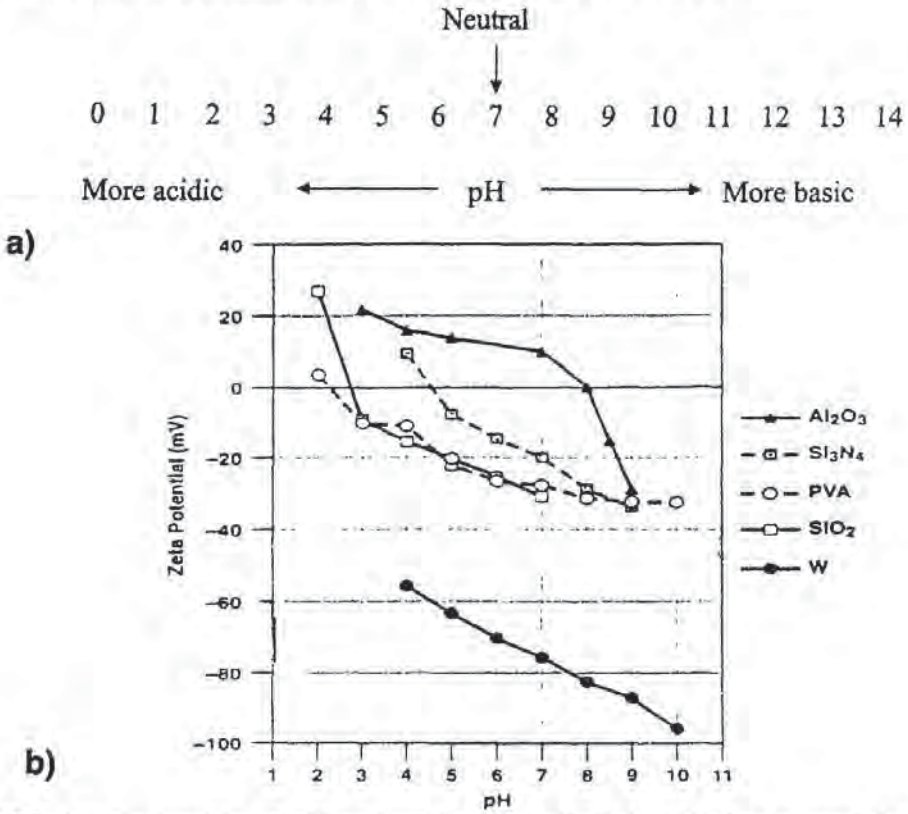


Fig. 8-75 a) Relationship of pH and acidity and alkalinity. b) Zeta potentials versus pH for particles of various types of slurry particles and materials in an aqueous solution.

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

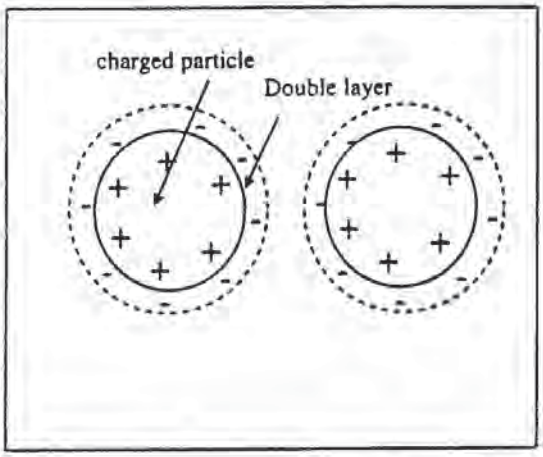


Fig. 8-76 Electrostatic repulsion keeps particles separated by repulsion of charged double layers.

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

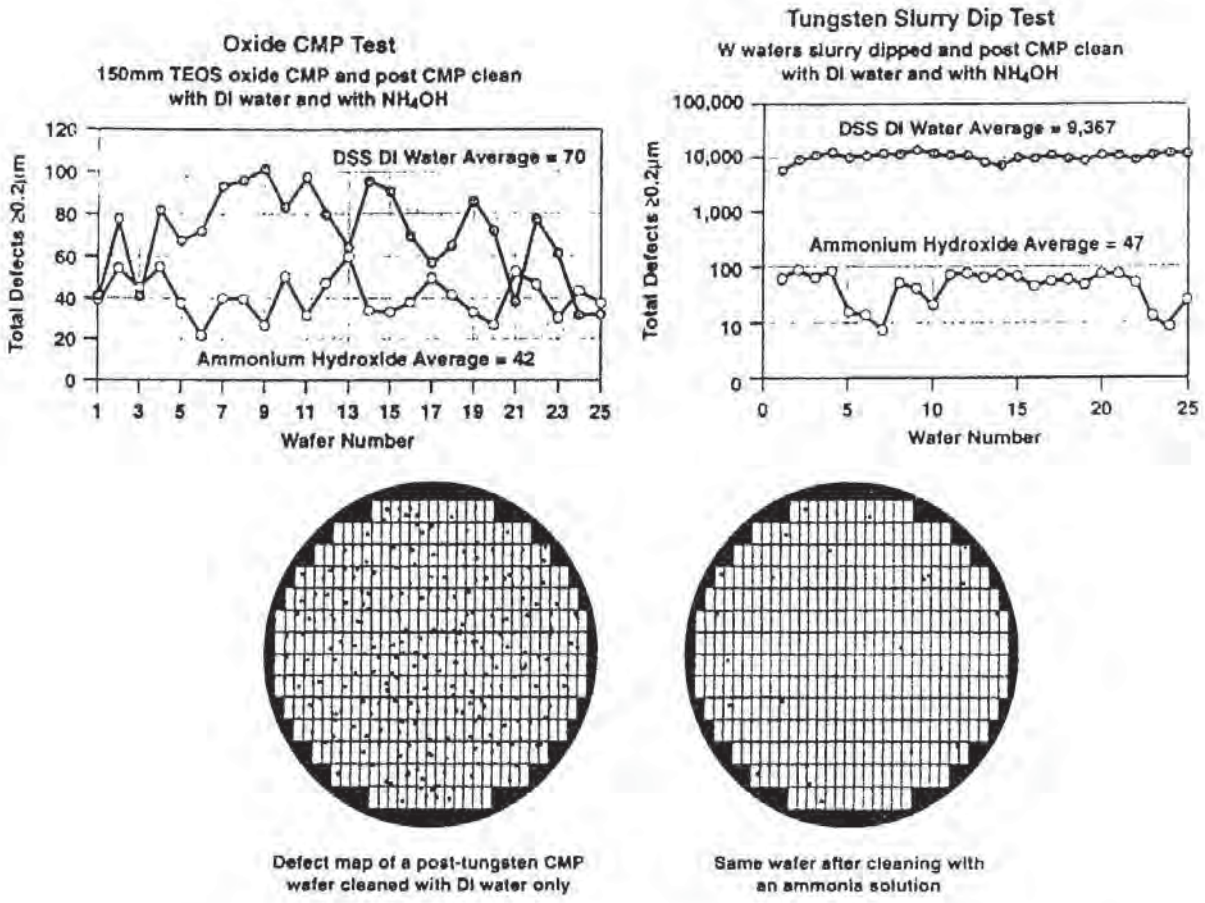


Fig. 8-77 Brush cleaning results comparing use of H_2O(DI) vs. NH_4OH: a) Oxide-CMP cleanup; b) W-CMP cleanup.

(*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]

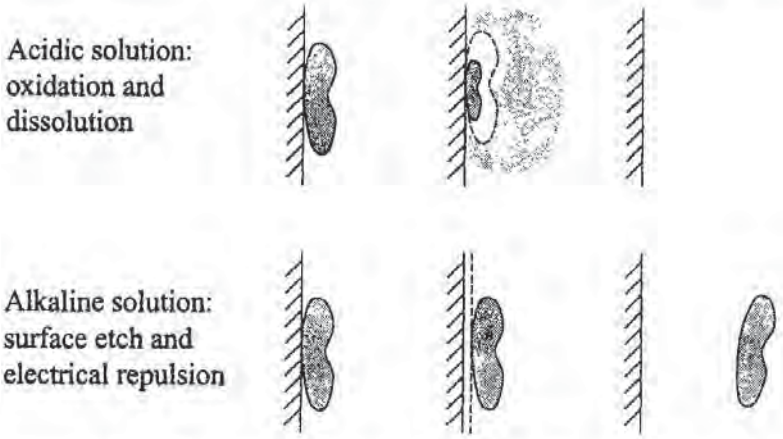


Fig. 8-78 Wet-chemical mechanisms that remove particles from surfaces: a) in a powerful oxidizing agent; b) in an alkaline solution. Courtesy of Semiconductor International.

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

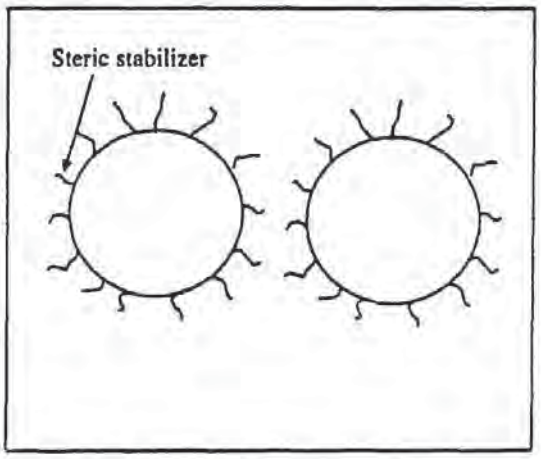


Fig. 8-79 Steric-stabilization of particles in aqueous solution. Particles are stabilized by attaching a molecule to the surface. The molecular layer causes an entropic repulsion as the particles approach each other.

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

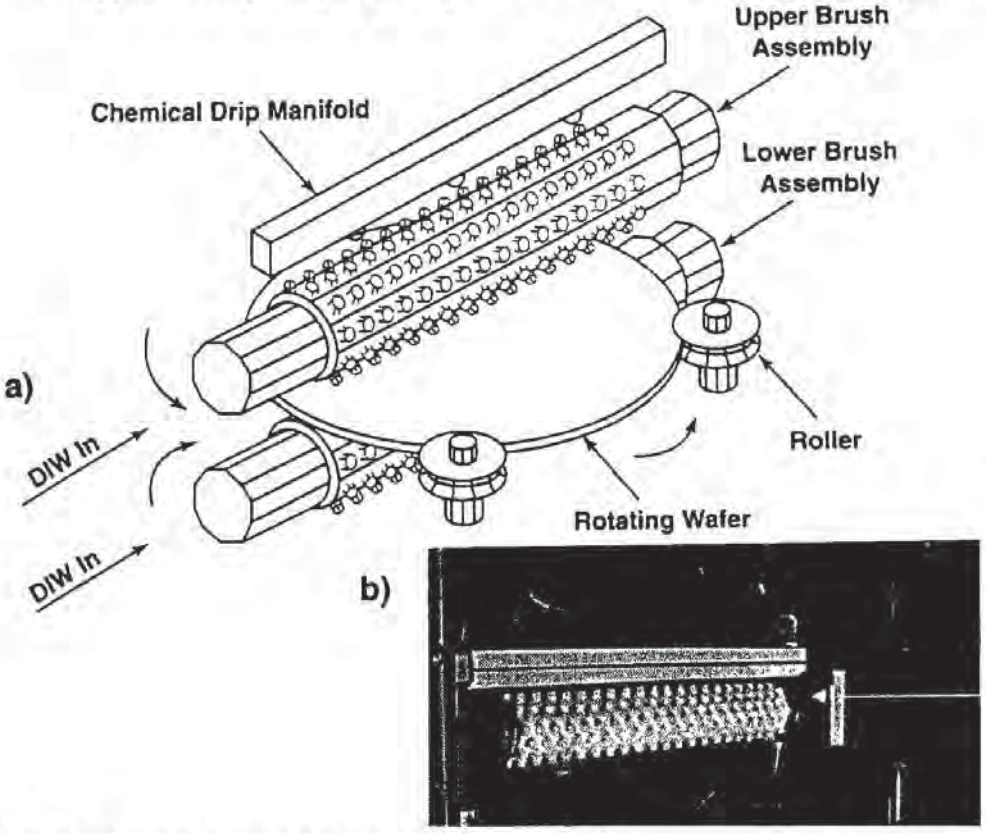


Fig. 8-80 a) Schematic of a double-sided brush scrubber.[65] Courtesy of Lam Research; b) Photograph of a PVA brush. Courtesy of Solid State Equipment.

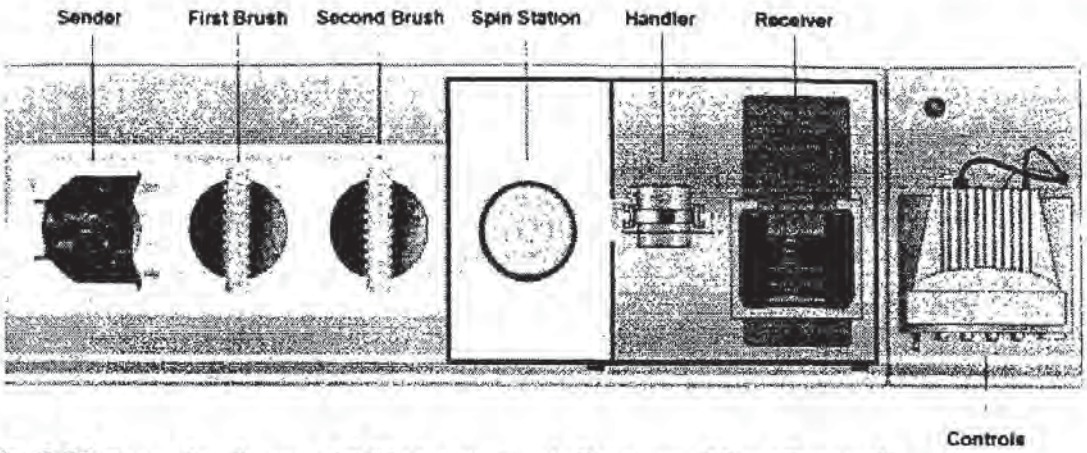


Fig. 8-81 Example of a post-CMP cleaning tool. Courtesy of Lam Research.

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.

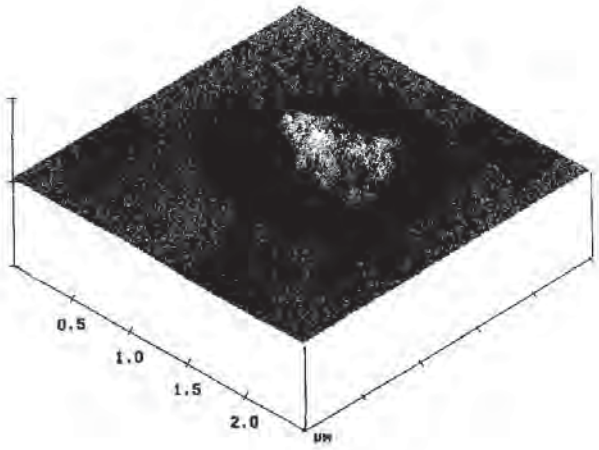


Fig. 8-82 AFM image of an agglomerated-silica slurry-particle in a contact hole.

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.

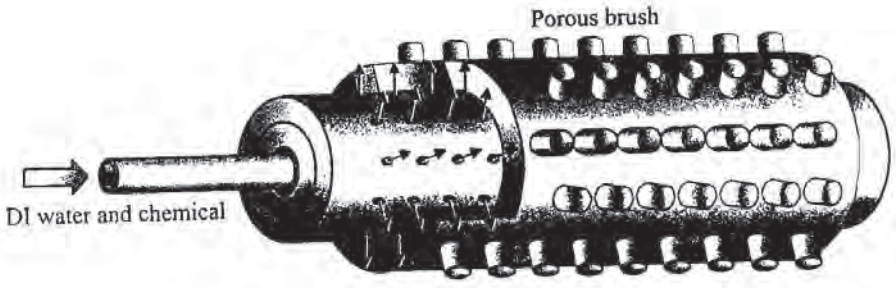


Fig. 8-83 *Through-the-brush* Chemical Delivery for post-CMP cleaning.

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).

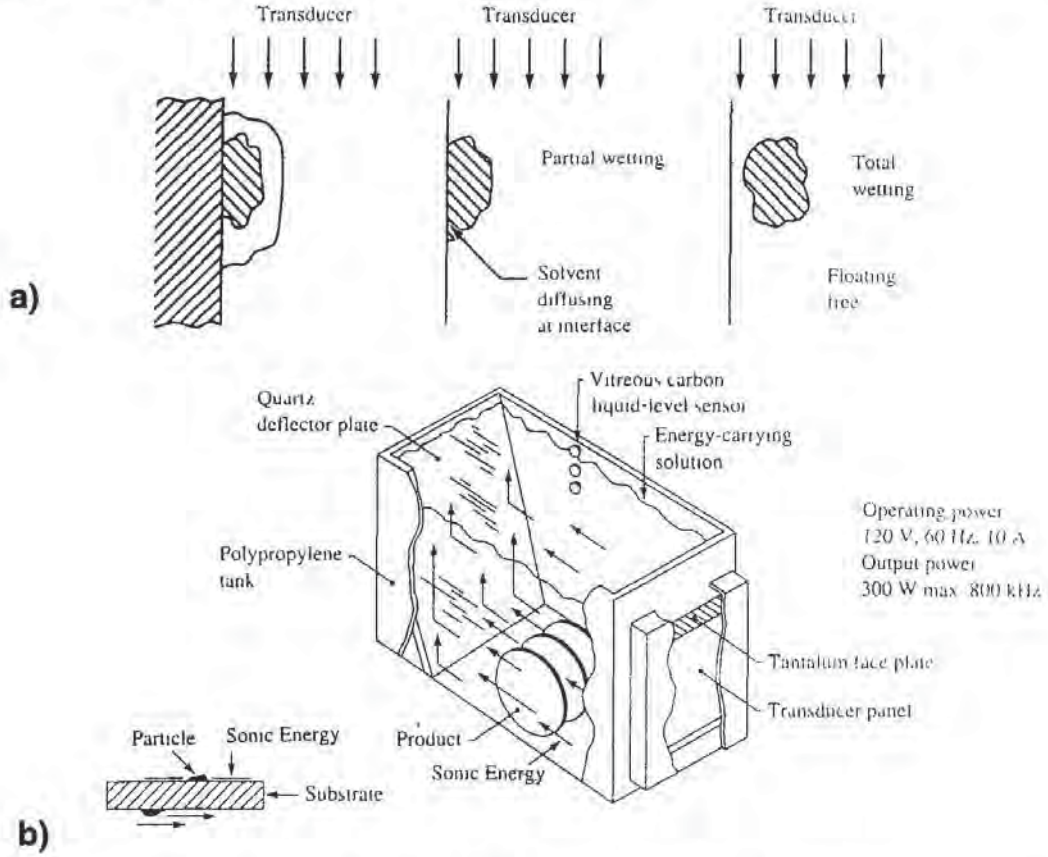


Fig. 8-84 a) Model for the mechanism of particle removal by megasonic scrubbing; b) Schematic of megasonic cleaning tank.

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.

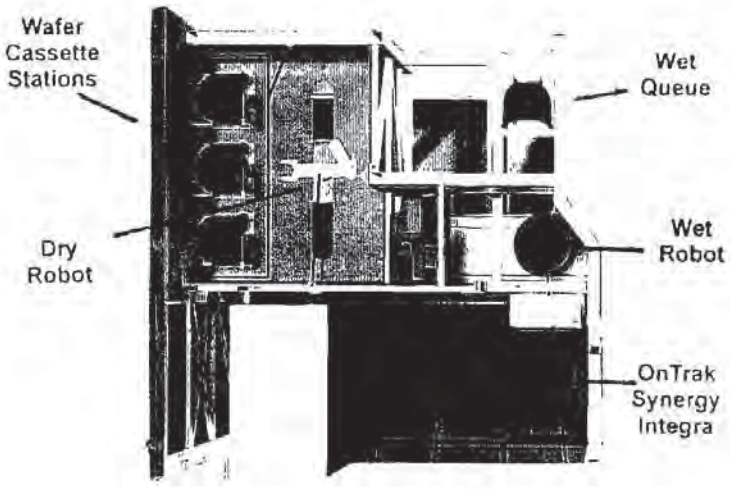


Fig. 8-85 Schematic of the Lam Research Synergy® post-CMP cleaning tool. Courtesy Lam Research.

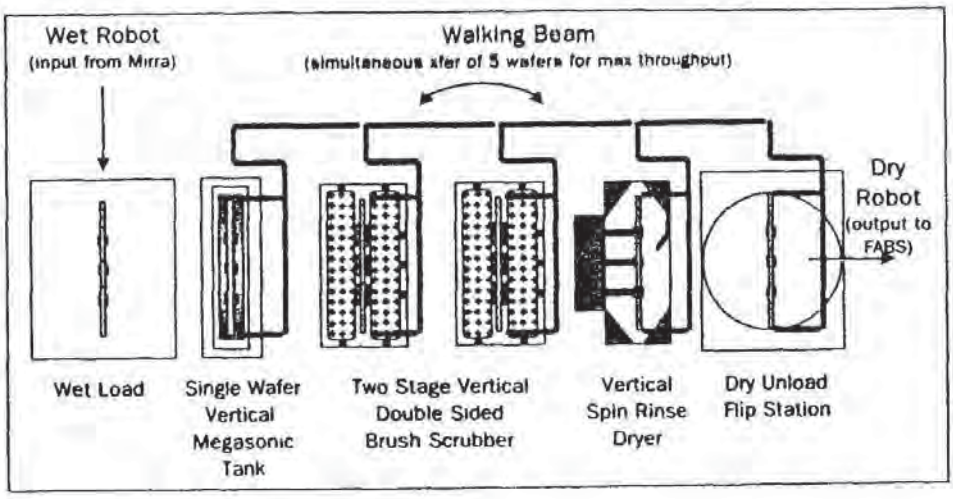


Fig. 8-86 Mirra Mesa® cleaning module layout. Courtesy of Applied Materials, Inc.

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

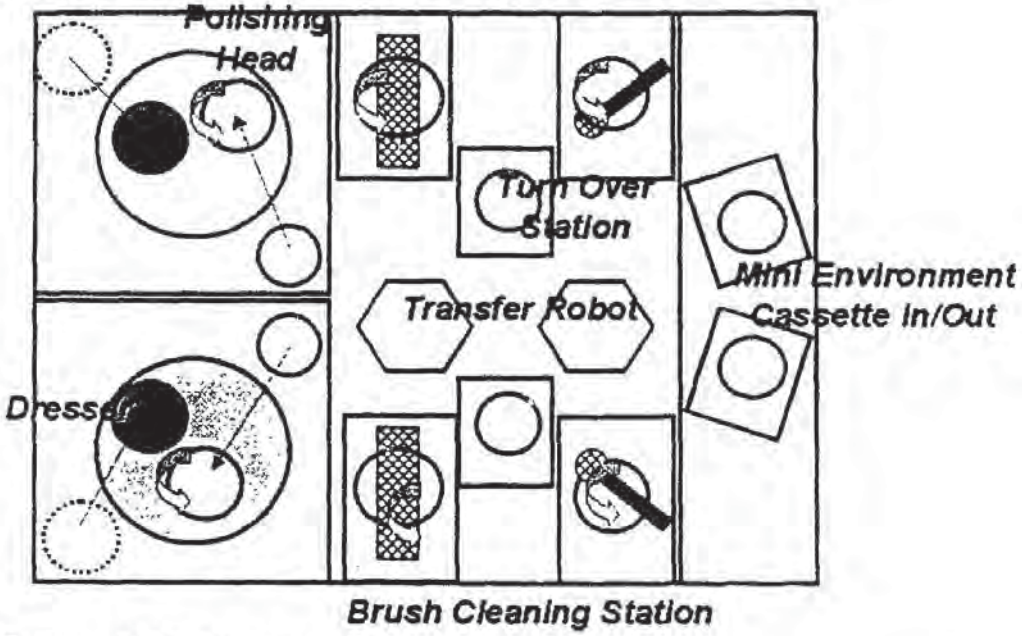


Fig. 8-87 Ebarra CMP cleaning module. Courtesy of Ebara.

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

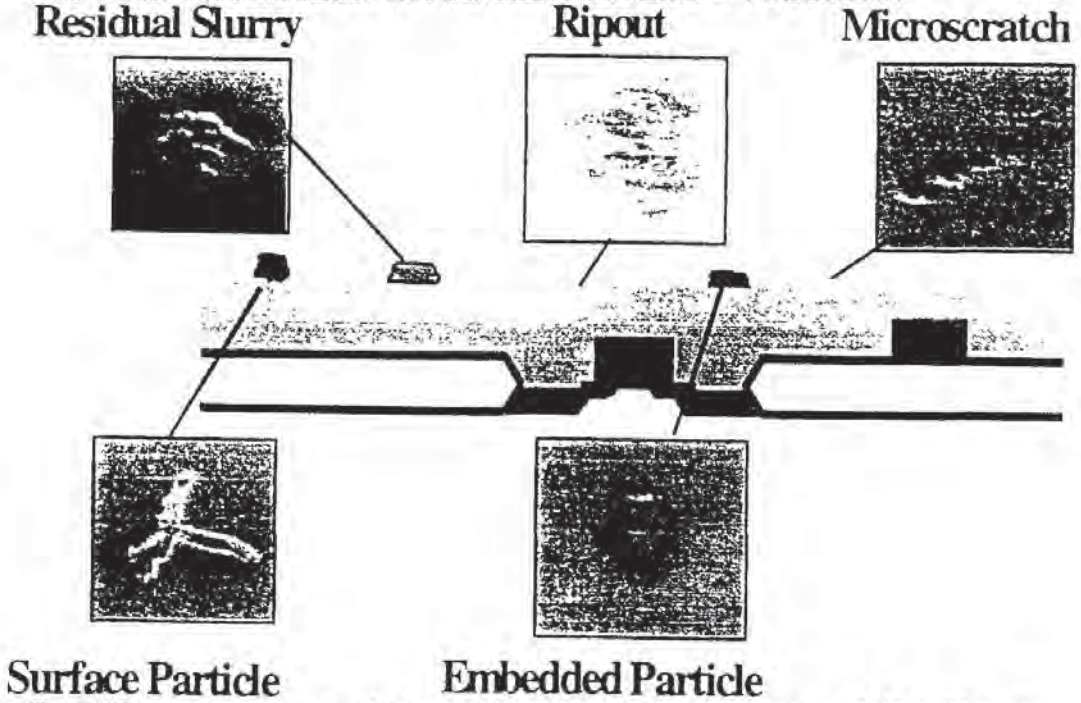


Fig. 8-88 Defects that arise during oxide CMP processes. Courtesy of KLA-Tencor.

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.

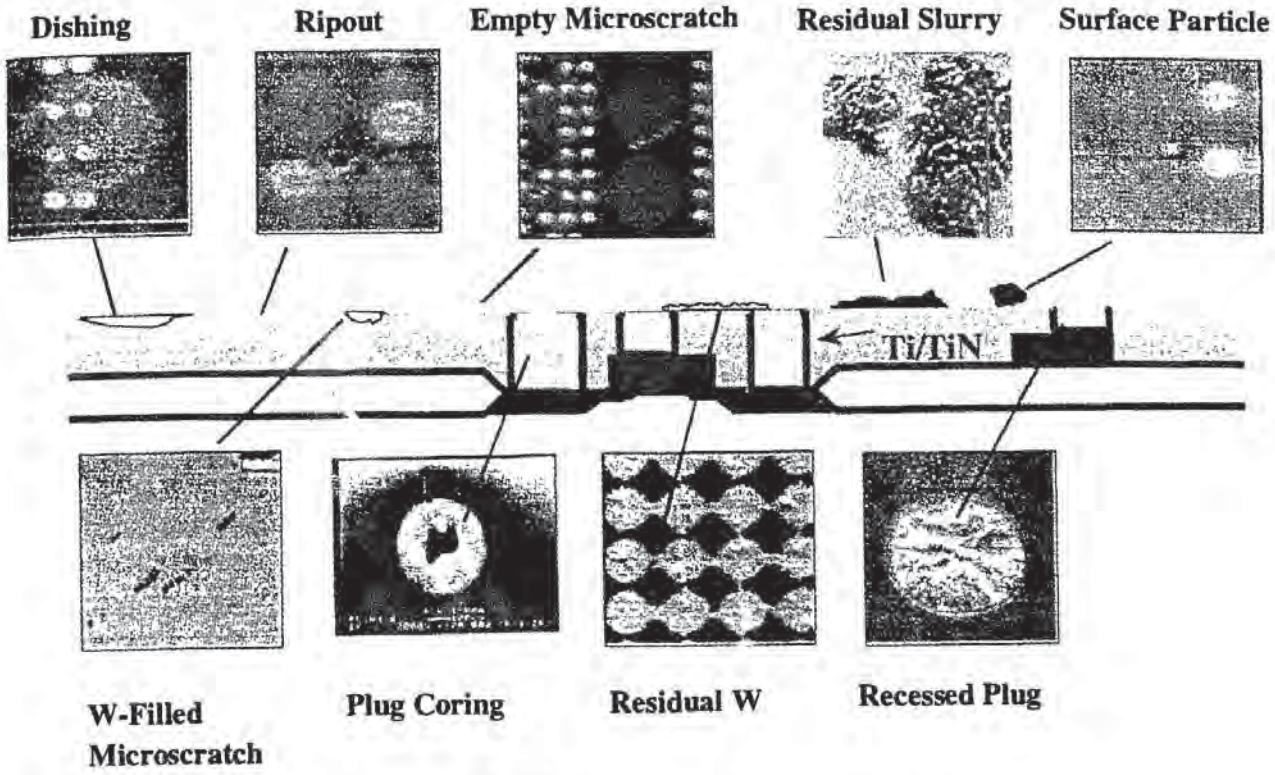


Fig. 8-89 Defects that arise during W-CMP processes. Courtesy of KLA-Tencor.

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

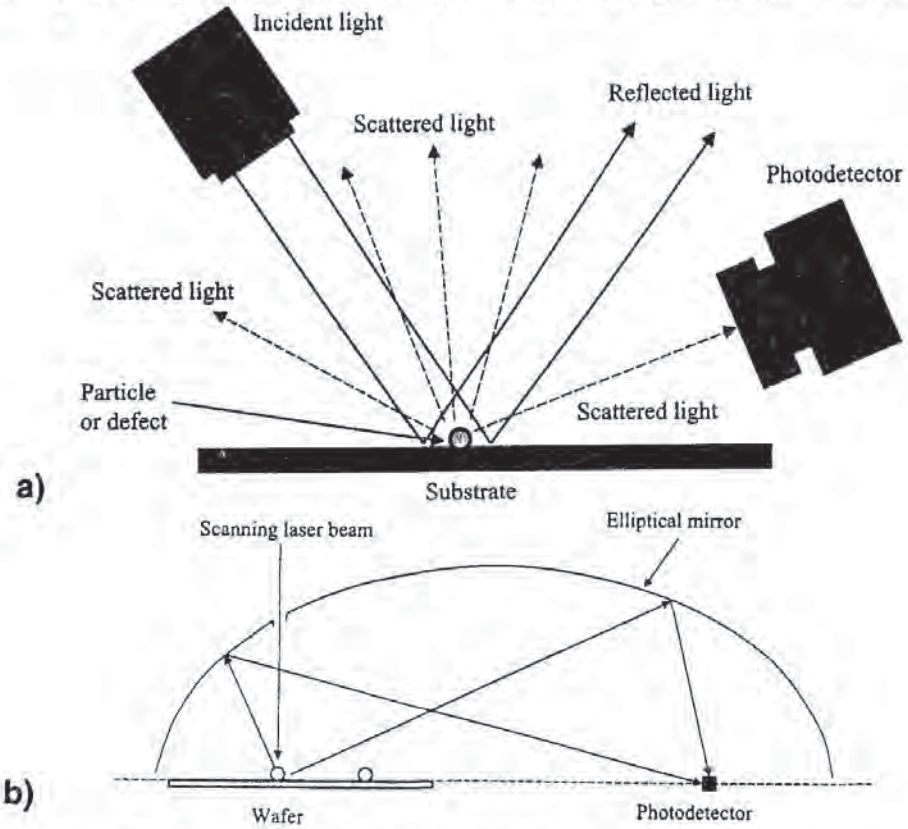


Fig. 8-90 a) Particle detection by Darkfield Light-Scattering; b) Light-scattering particle detector with an elliptical mirror.

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).

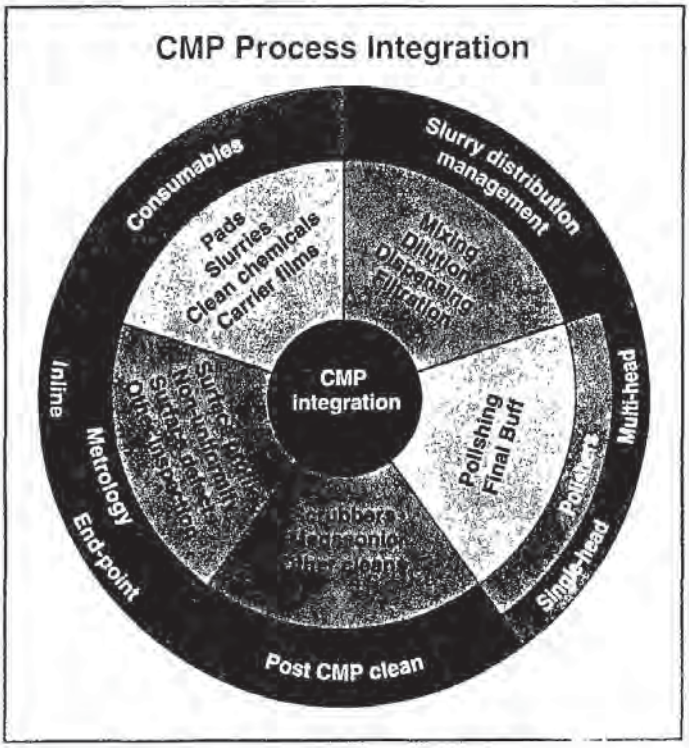


Fig. 8-91 Many components of CMP must be successfully integrated to develop a production-worthy CMP process. Reprinted with permission of Semiconductor International.

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

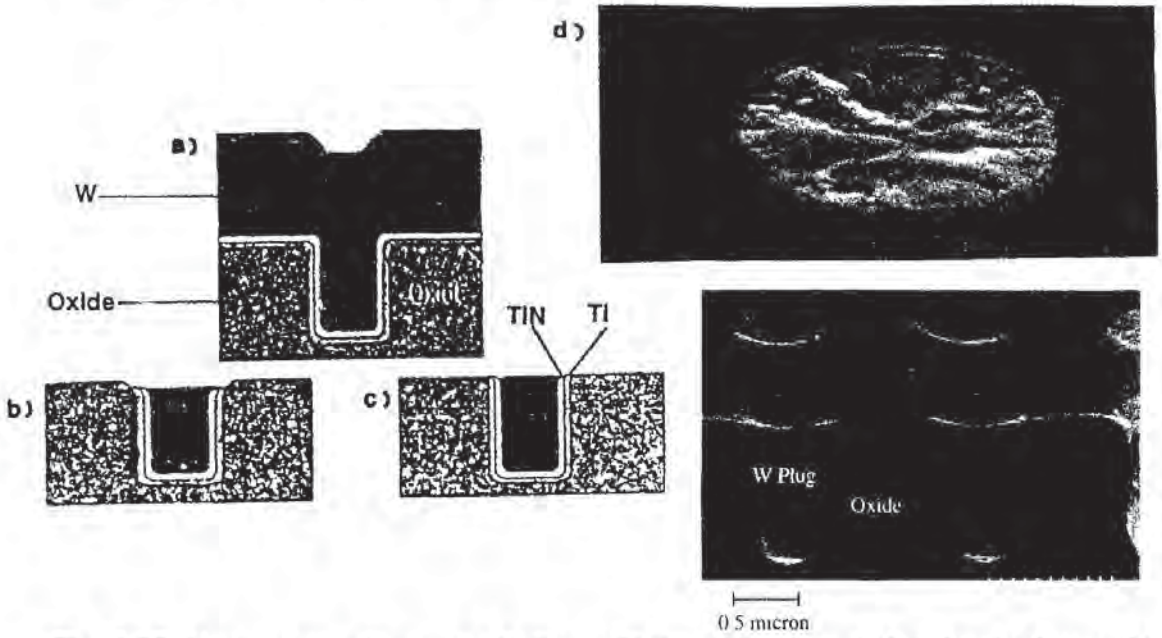


Fig. 8-92 W-plug formation using CMP: a) W deposition over TN/Ti barrier layers; b) CMP of W and TiN/Ti leaves recessed plugs; c) Oxide buff eliminates recess.[84] Reprinted with permission of Solid State Technology. Published by PennWell.; d) SEMs of W-plugs after W and TiN/Ti CMP steps and oxide buff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

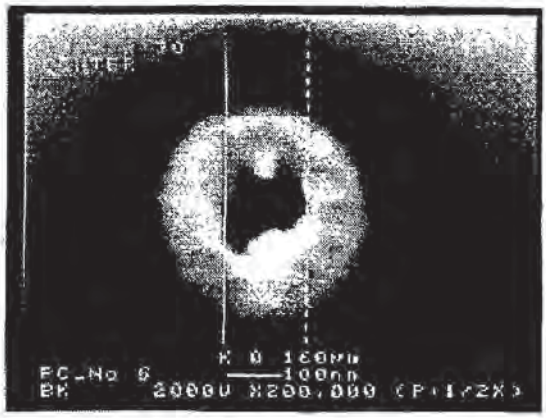


Fig. 8-93 Plug-coring.

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

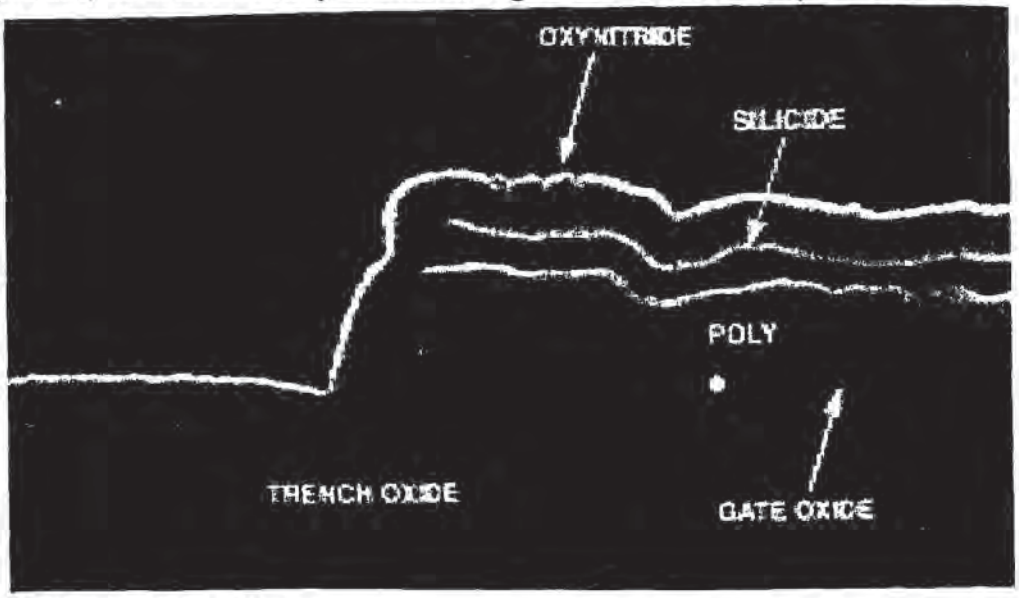


Fig. 8-94 CMP of polysilicon.

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,

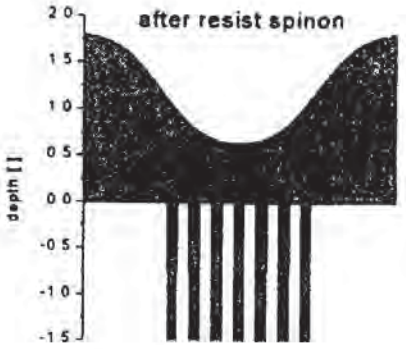


Fig. 8-95 CMP of photoresist.

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

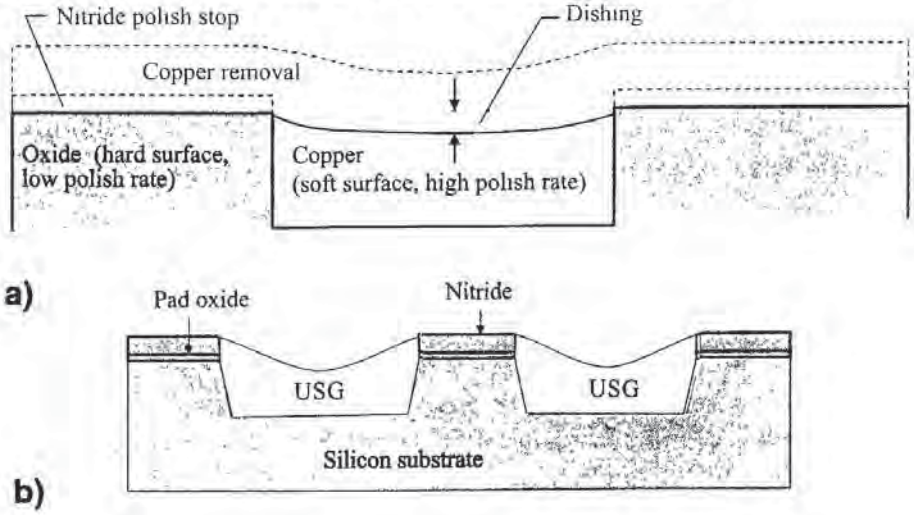


Fig. 8-96 Dishing-effect, shown here for: a) CMP of C in oxide-trenches; b) CMP of oxide in Si-trenches during formation of STI.

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).

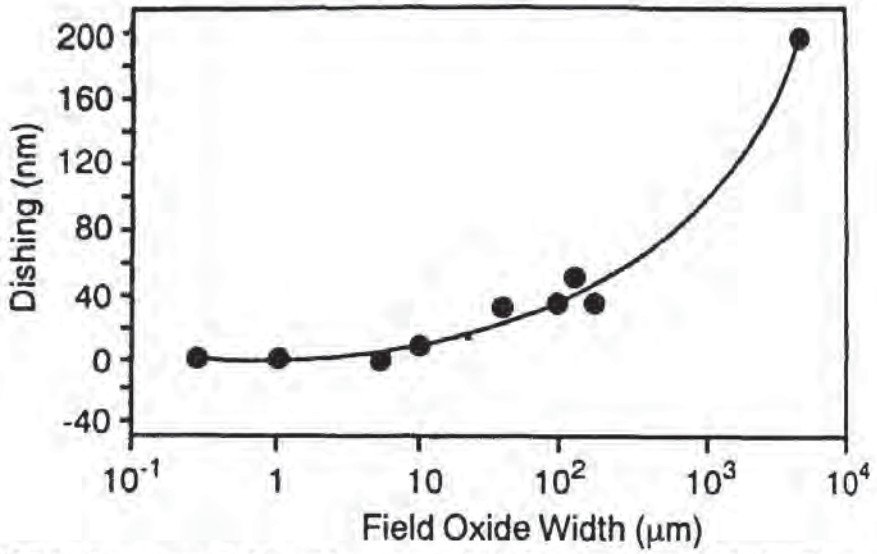


Fig. 8-97 Dishing versus field-oxide-region width for a specific oxide CMP process.

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

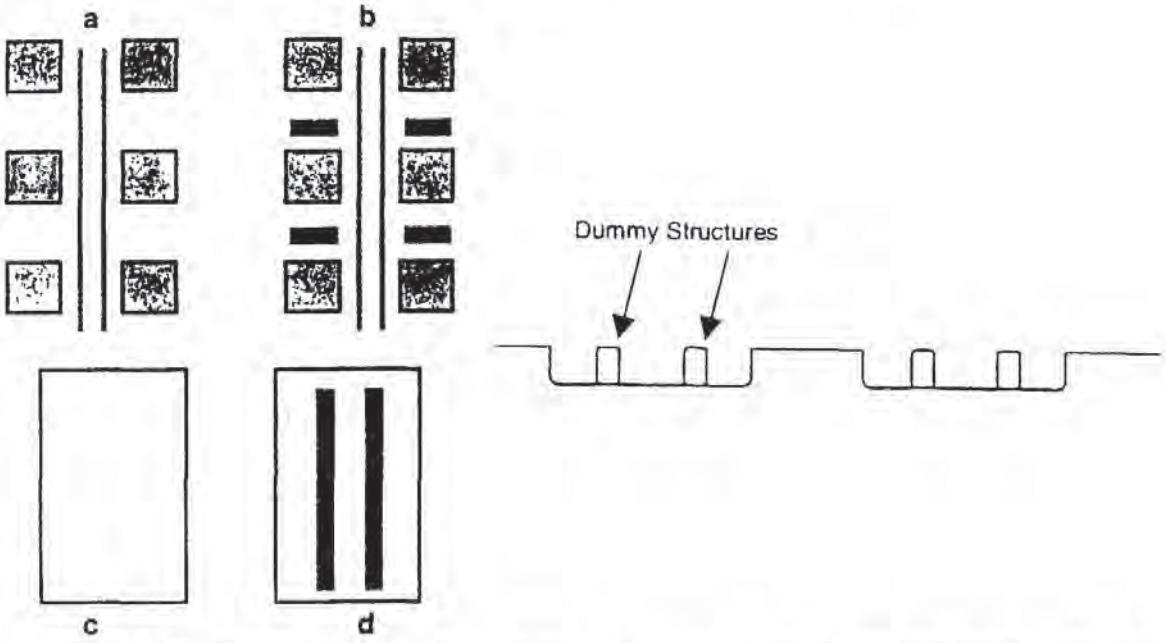


Fig. 8-98 Dummy (b) and dishing support features of small (a) and large wafer (c) features can prevent dishing.[89] e) Cross-section of dummy features in a trench.

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

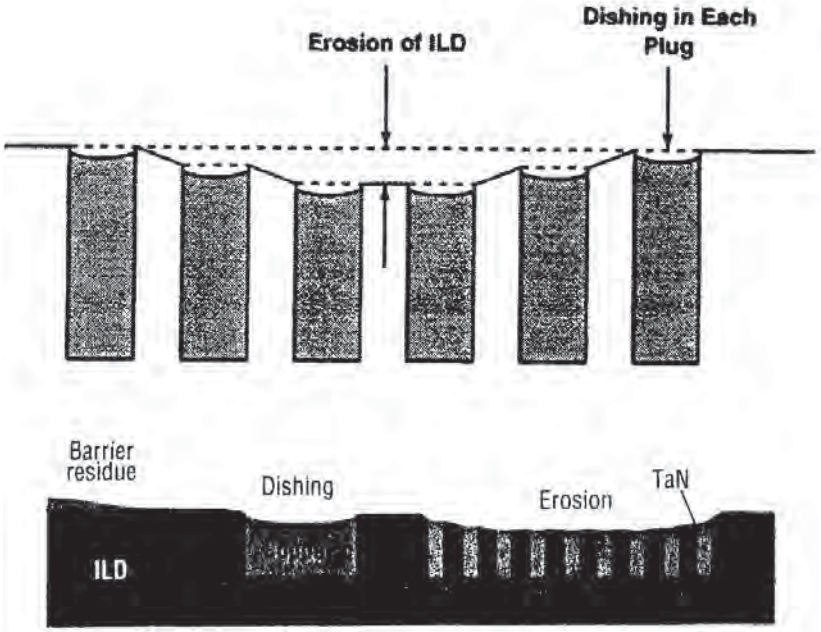


Fig. 8-99 Erosion of the ILD that occurs during over-polish of Cu during CMP.

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,

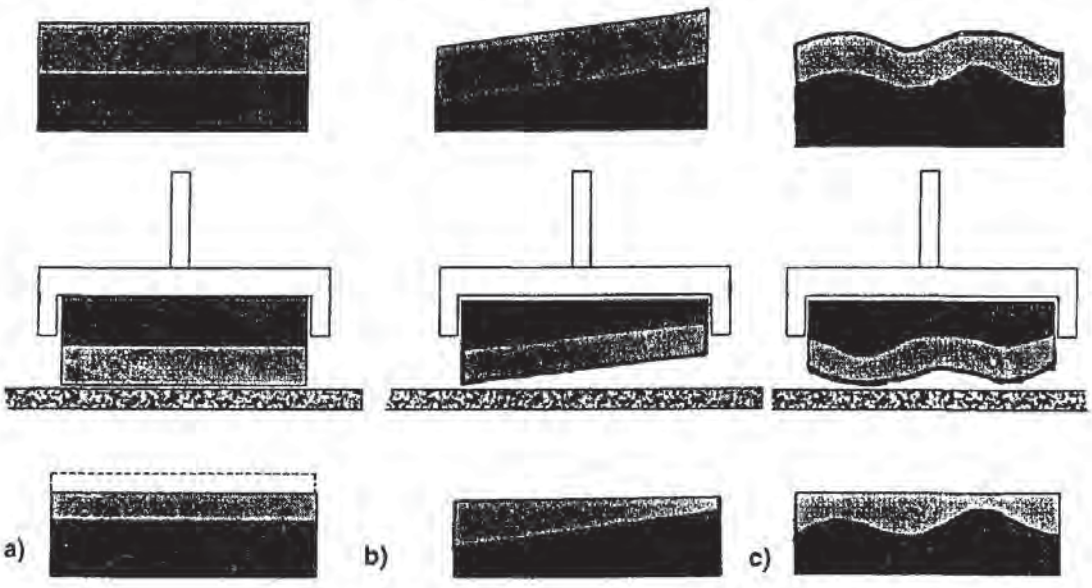


Fig. 8-100 If a rigid pad is used for CMP: a) A film on a flat-wafer is uniformly polished; b) A film on a tapered-wafer is polished thinner over the thicker end of the wafer; (c) For wafers with variations in wafer-thickness across the wafer, the film is polished thinner over the thicker wafer regions.

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.