Identifying front-end challenges for 90nm design

By Gregg Higashi
CTO and Director
Emerging Technologies
Front-end Products
Business Group

Thorsten Lill
Technology Director
Silicon Etch Product Unit

Applied Materials Inc.

Challenges in the front-end of line (FEOL) for the 90nm technology node are similar to the challenges faced in previous nodes. Transistors are being scaled, and one or two new materials and processes are being introduced. Although densities are being improved, it is becoming increasingly difficult to achieve the performance improvement expected for each successive node. Sufficiently high drive currents with low off-state leakages supporting the requisite manufacturing control are particularly challenging.

The potential solutions being attempted at various sites around the world vary but a number of trends are clear. Innovative new processes, process sequences and control are even more critical to achieve aggressive 90nm scaling targets. Partnerships between IC manufacturers and tool suppliers are crucial and early engagement is the only way to meet the required timelines. Understanding processes, process interactions, devices and device technologies help limit the number of potential paths pursued and keeps development schedules on track.

FEOL scaling is about density improvement while simultaneously improving transistor performance. Thus, one can divide scaling into two parts: isolation and transistor.

Isolation scaling

At sub-90nm device nodes, implementation of shallow trench isolation (STI) becomes more challenging, incorporating aspects of trench definition (lithography, etch), liner oxidation, trench fill with deposited oxide, CMP and another thermal oxidation to grow the sacrificial oxide. The major challenges with scaling STI have to do with scaling aspect ratios and with the control of the corner rounding and the stress of the STI on the active area of the transistor. The STI aspect ratio (depth of trench/width of trench) is estimated to be about 3.5:1. The challenge for etch is to precisely control the 400nm depth simultaneously with the taper.

Perhaps a more difficult challenge is to fill the trench after the etch without leaving keyhole like voids behind due to pinch off of the trenches during the fill. High-density plasma (HDP) enhanced CVD films have been found to be well-suited to this task. HDP CVD SiO_2 films have become the industry standard for this process. The impact of STI stress is also a difficult challenge. The stress is particularly troublesome for two reasons: it has been found to cause NMOS drive current degradations on the order of 10 percent in narrow active area devices, and has also been found to cause defects and device leakage.

The liner oxidation rounds the top and bottom trench corners. Top corner rounding is critical to avoid formation of parasitic corner transistors, to minimize shifts in threshold voltage and to prevent premature gate dielectric breakdown. Bottom corner rounding minimizes formation of stress-induced defects that can result in junction leakage when propagated in subsequent steps. The sacrificial oxide, following polishing as well as pad oxide and nitride removal, minimizes stress that can lead to defects and device yield degradation.

We have found that Applied Materials' patented in-situ steam generation (ISSG) addresses the limitations of other thermal oxidation methods for growing the STI liner and sacrificial oxide. This approach creates atomic oxygen radicals that round the top and bottom corners of the STI trench in a single step with minimal consumption of active area and with improved conformality due to reduced crystallographic-orientation dependence. These physical properties result in improved device yield for memory and logic devices.

Using this technique also improves device yield when used for sacrificial oxidation due to a lack of trench reoxidation at the top corner and edge. There is minimal diffusion of oxygen to the trench corners and sidewalls due to RTP's fast temperature ramp rates and fast ISSG oxide growth rates, thereby minimizing subsequent stress-induced silicon defects. Our approach is being adopted for production of STI liner oxidation and sacrificial oxidation for sub-90nm devices.

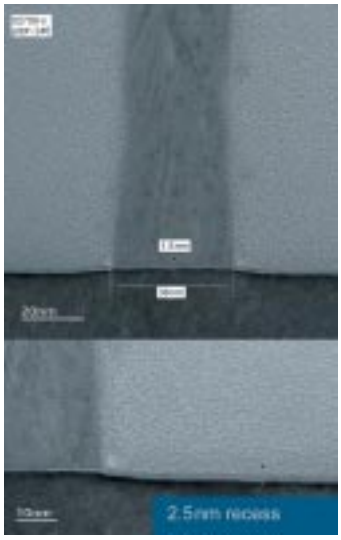

Figure 1: (a) Etched poly gate stack showing 90nm gate length capability; (b) perfect etch stop demonstrated at gate edge.

Another way to control or offset the stress is to add films to oppose the stress of the STI. A common practice for 130nm technology is to add a nitride layer in tensile stress just prior to the first interlayer dielectric. Drive current improvements of 10 percent have been reported. Other techniques under consideration include selectively grown epitaxial SiGe, more commonly used in hetero-junction bipolar transistors. SiGe is a new material to CMOS and has a number of other properties besides adding strain to the system.

Transistor scaling

There are two broad considerations in scaling transistors: performance and reliability. Performance is largely determined by scaling three parameters: gate length, gate oxide thickness and junction depth. After scaling these parameters, companies need to ascertain that the devices still function reliably. The industry standard is to ask that the devices be projected to last at least 10 years under normal operating conditions.

Gate length control

The main challenge in gate etching remains the control of the critical dimension (CD). With the 90nm technology node in pilot production and 65nm in development, the CD control requirements are becoming more stringent. Typically the physical gate length for the 130nm and below technology nodes is 50 percent of the half pitch. For the 90nm node we look for instance at physical gate lengths of 45nm (**Figure 1**). The three sigma CD uniformity requirement is usually given as a percent budget of the physical CD, in most cases 10 percent. This means that the post etch CD control requirement including photolithography is about 5nm for the 90nm and 3nm for the 65nm technology nodes. Apart from the CD control, the reduction of "gate oxide recess" is a critical requirement. A gate oxide recess of 3nm and below is considered acceptable (**Figure 1b**).

It is mandatory to compensate for the incoming litho CD uniformity to achieve this level of post etch CD uniformity. An advanced gate etch system has to address systematic CD non-uniformities on the incoming wafers. The main contributors to the CD non-uniformity of the incoming wafers are the within die proximity effect (dense/iso effect), systematic across wafer effects (side to side or center to edge), and wafer-to-wafer (WTW) and lot-to-lot. The relative contribution from these sources can vary from fab to fab depending on the tool setup and integration scheme (use of BARC, hardmask etc.). The etch system and process have their systematic CD non-uniformities on all three levels as well. One etch specific CD effect is the doping effect; the

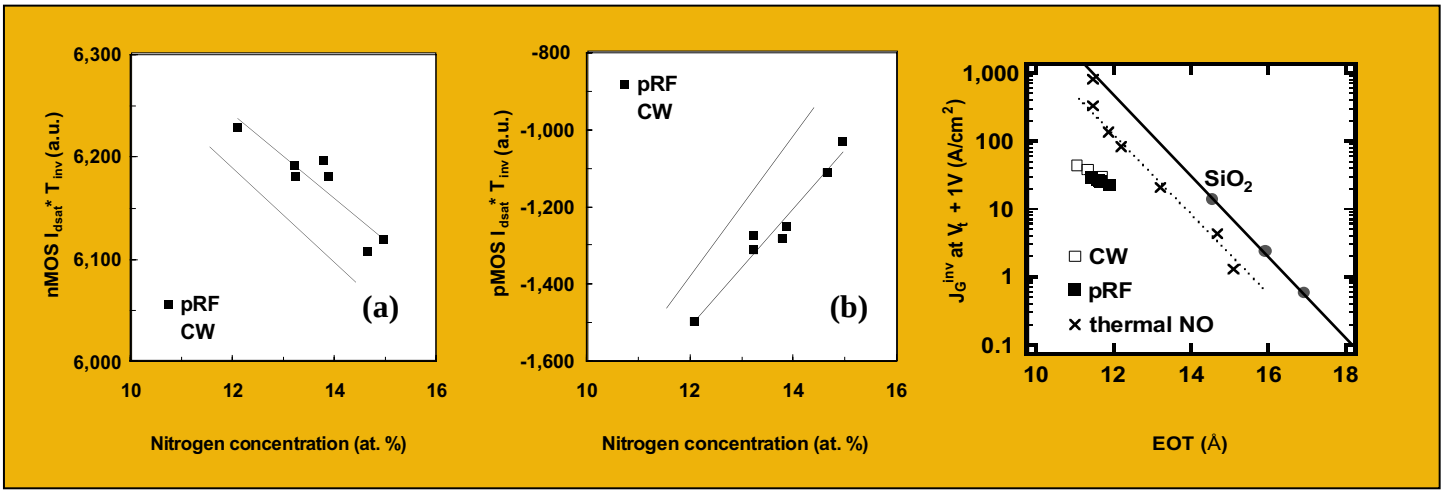

Figure 2: (a) nMOS gate leakage as a function of EOT for CW and pRF plasma and thermal NO oxynitrides. At fixed EOT, pRF splits have less leakage current than CW splits; (b) Normalized saturation drain current for long channel nMOS and pMOS transistors.

n-doped lines tend to loose more line width than the undoped or p-doped lines.

CD control across die is driven by the choice of the physical and chemical plasma properties. Since anisotropic etching is achieved by creating the proper balance between etch and deposition, the proximity effect for both processes will define the overall CD microloading for a given step. One good example how this can be used is the addition of C_xF_y containing gases to a poly-Si gate main etch to change sidewall passivation mechanism from SiO_xCl_y based to C_xF_y based with the effect of reducing the dense–iso profile and CD difference. It appears that the gas phase originated carbon sidewall shows less proximity sensitivity than the silicon oxide based passivation, which is created primarily by redeposition from the etch front and is therefore sensitive to the open area next to the feature. It is worthwhile to mention that the addition of CF_4 or NF_3 also reduce the doping effect quite significantly.

Frequently the integration requires an overall CD reduction from a printed CD to a final physical CD of the gate line. This is usually accomplished in a process step called resist "trim". Because of its more isotropic nature, this step can be designed to be "isolated trim fast" or "dense trim fast". Similar to the silicon gate etch, the main technique for the tuning of this step is to vary the balance between etch and deposition processes. Bias power and the addition of C_xF_y gases can be used as parameters to adjust the dense–iso etch behavior. One major concern for the trim step besides CD uniformity and CD microloading is the available resist budget, especially if there is a subsequent mask open step involved. We have found that using an advanced patterning film consisting of a carbon based material with an optional dielectric cap layer as a hardmask film offers the resist thickness after trim to allow the opening of the thin dielectric top layer. Subsequently, the resist can be stripped and the dielectric layer can be used to etch the carbon-based hardmask. The dielectric layer is removed easily during the poly-Si main etch. After the complete gate etch, the carbon hardmask can be stripped in-situ. This constitutes another very compelling feature of this mask stack since the strippability of dielectric SiON-based hardmasks is an issue especially for highly doped and unannealed poly-Si gates due to attack by hot phosphoric acid.

CD distribution across the wafer depend on the scale of the wafer size and are addressed by the design of the plasma etch chambers and its tuning features. At least three parameters can be used for center to edge CD distribution control in the silicon etch chamber: ion flux density, neutral flux density and wafer surface temperature.

WTW and lot-to-lot CD control require a stable and repeatable process chamber. This includes well-defined chamber wall conditions since this is where a large fraction of the radical recombination occurs. We have found the combination of a self-clean process portfolio and waferless dry cleans address this issue. The addition of fluorinated gases suppress the formation of SiO_2-based deposits on the chamber walls and reduce the dry clean time significantly which improves the etch productivity. The use of the waferless dry clean allows to run non-clean, i.e. depositing processes in combination with self-clean processes in mixed mode. If the waferless dry clean is controlled by endpoint, the mixed production can be automated.

A stable chamber requires close monitoring and control of all chamber parameters via fast data acquisition. Full spectrum emission as well as plasma parameters like ion energy and flux to the cathode also need to be monitored and analyzed. Data reduction techniques such as neural networking and principal component analysis are used to realize fault detection and real-time process adjustments. Process adjustments can also be made based in the information from the incoming wafer or a previously etched wafer. This information can stem from on-board or standalone metrology tools. On-board metrology solutions offer dramatic reduction in cycle time and are essentially the only way to make process adjustments based on results from an etched wafer (feedback control). One of the first applications to benefit from this technique is adjustment of the resist trim time based on pre-etch resist profile measurements (feed forward control). This integrated technology is now in production and provides valuable insight in the correlation between lithography and etch on a level that can be only achieved when every wafer is measured. To provide the option to adjust the trim time based on the outgoing wafer (feedback control) we found an integrated wafer treatment allows measuring the profile of the etched wafer before it leaves the etch system, i.e. before clean. This technique provides further reduction in WTW, lot-to-lot CD uniformity and a dramatic reduction in cycle time since send-ahead lot adjustments of the trim time can be automated.

These examples show that advanced gate etching is relying increasingly on a wide range of technologies to address the ultimate goal of tightest possible CD control for each transistor on each manufactured chip.

Gate oxide scaling

Gate oxide thickness scaling is critical to achieving device performance targets, but scaled dielectrics must first meet leakage and reliability requirements. Starting with the thicknesses and thermal budgets required for the 130nm node, ~2.0nm and ~105°C for 1.6s respectively, nitrogen incorporation in the gate oxide has been a prerequisite for stopping B-penetration from the gate poly into the Si substrate. The nitrogen content of the dielectric provides the additional benefit of reducing gate leakage. Oxynitrides have been successfully extended to 90nm technologies with thicknesses ranging from 1.2–1.6nm using Decoupled Plasma Nitridation (DPN) technology clustered with an in situ RTP anneal. This in situ anneal has been found to be particularly important for im-

proved WTW and within wafer (WIW) uniformity. This improved WTW performance is critical for production control of the process. Queue time between production steps and adsorbed contamination on the ultrathin gate dielectrics has also been shown to affect final EOT by more than 1A.

The main challenge for extending oxynitride gate dielectrics to 65nm is meeting the leakage and reliability requirements at 1.0nm without a significant degradation in mobility. Fully clustered (base oxide through PNA) gates using DPN technology have been shown to achieve sub 1.2nm EOTs with acceptable leakage and reliability, while maintaining high mobility (**Figure 2a**). The clustering is shown to improve EOT scaling, WTW and WIW uniformity as well as transistor performance. Hardware improvements such as pulsed RF technology is shown to minimize V^{th} shift and improve drive current by lowering the electron temperature (kT_e) of the plasma and control of the nitrogen profile in the oxide (**Figure 2b**).

Ultra-shallow junction scaling

Scaling transistor junctions is a complex problem requiring simultaneously optimizing junction depth, sheet resistance and lateral abruptness. Each of these parameters plays a critical role in determining the transistor's short channel performance. These parameters are typically controlled using ion implantation and rapid thermal annealing technologies. The discussion of selectively grown epitaxial Si and SiGe raised source/drain (S/D) and S/D extensions will follow.

The International Technology Roadmap for Semiconductors has specified a 90nm junction depth between 15nm to 25nm with a sheet resistance $< 660\Omega^2$. These values are comfortably met with conventional implant and spike anneal (1.7s residence time within 5°C of Tpeak) technologies. Hence, the main focus of equipment suppliers has been productivity improvement and process controllability for 90nm. High productivity in ion implantation is most difficult for B because of the ultra-low energies required to meet the ultra-shallow junction requirements. 500eV B implants at doses approaching $1E15/cm^2$ are commonly used for S/D extensions. We have achieved high productivity and precise repeatability has been achieved using advanced deceleration lens technology and beamline designs along using moderate deceleration. Thirty-five wafers per hour are now typically processed. The importance of achieving good junction abruptness also needs to be highlighted. Simple implanted and spike annealed junctions result in an abruptness of approximately 8nm/decade of depth into the Si wafer. This can be improved to 6nm/decade by first amorphizing the Si prior to implant using a Ge implant. Further improvement to 5nm/decade can be achieved by co-implanting a species such as F that moderates B diffusion. Although this junction abruptness is slightly higher than the ITRS specification of 4nm/decade it appears that 90nm devices are functioning adequately.

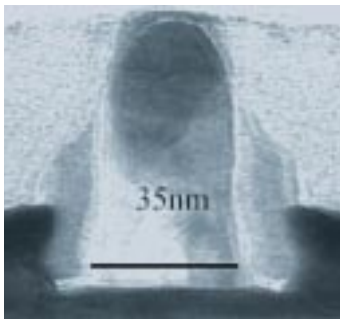

Figure 3: Ultrathin SOI under the gate with thick elevated source/drains adjacent to the gate that are later silicided with Ni.

The key to successful high volume manufacturing is, of course, the level of control that can be achieved with these aggressively scaled junctions. Implant dose and energy accuracy are important, but research has shown that spike anneal temperature uniformity is the most critical parameter to be controlled. The spike temperature uniformity has a strong effect on device performance, because the lateral position of the S/Dextensions and the resulting effective channel length are exponentially dependent on the temperature of this anneal. This temperature sensitivity can been seen in the pMOS transistor threshold voltage which varies by more than 2mV/°C, leading to a temperature control requirement of < 5°C for 90nm technology. This sensitivity increases as device dimensions are scaled to the 65nm node.

These challenging device requirements are being met with spike anneal capability that controls temperature uniformity at all points on all wafers within 5°C, 3-sigma. This tight performance improves device yield and enables more of the highest-speed chips to be produced per wafer.

Epitaxial Si, SiGe

As mentioned previously, stress is an important parameter to control in optimizing device performance. S. Thompson, et. al. have reported using selective SiGe in the S/D areas results in a significant improvement (up to 20 percent) in the drive current of MOSFET devices. In addition to mobility improvements associated with strain, there are other important benefits of using SiGe. First, SiGe has a smaller band gap than that of Si and, consequently, a reduced Shottky barrier at the semiconductor-silicide interface. Second, Ge increases incorporation of dopants into Si. These two factors contribute to reducing S/D contact and sheet resistances, increasing the drive current and the speed of a MOSFET device.

There are significant challenges in SiGe deposition. Not only film thickness but also the Ge and dopant concentrations should be uniform across the wafer and reproducible run-to-run. The necessity of controlling both thickness and two concentrations imposes a challenge by itself. Preparation of the Si surface before epitaxial deposition is important as any residual contamination or damage left by etch steps impacts negatively the quality of the Epi film or may result in no growth at all. In addition to the uniformity, reproducibility and surface preparation requirements, SiGe deposition should be selective, i.e., it should take place only on Si moats, with no deposition on dielectric areas.

As devices continue to scale and junctions become shallower, leakage can significantly increase. One way to prevent this from happening is to reduce the depth of the silicided area below the level of the gate dielectric. However, a thickness reduction of the silicide results in an undesirable increase of the sheet resistance. To address the leakage issue without increasing the contact resistance, one can use selective epitaxy to form elevated S/D areas above the level of the gate dielectric. Si Epi is used as a sacrificial layer as it is consumed by the silicidation process. Using elevated S/D becomes an absolute necessity in fully-depleted silicon-on-insulator devices. An example of such a transistor structure with elevated S/D is shown in **Figure 3**. The process flow involves fabricating spacers, performing Epi deposition, and metallization of the elevated areas.

The choice of metal for silicidation becomes critical in small-geometry devices. Among the factors to consider are Si (or SiGe) consumption during silicidation; thermal budget required to reach the low resistivity phase; low resistivity; thermal stability of the silicide. One of the most promising can didates is Ni that forms a low-resistance mono-silicide and mono-germanosilicide compounds.

The challenges in the FEOL for the 90nm technology node were rather modest. Extension of or upgrades to the existing equipment set were in large part sufficient to meet the performance and reliability targets required. For these processes, we demonstrated that advanced process control, uniformity improvement, productivity improvement, or focused process development were sufficient. Our new materials or processes introduced include plasma nitridation of gate oxides (DPN), carbon based hard masks and selectively grown Si and/or SiGe. Future technologies will be considerably more difficult. Presently being discussed are high-k gate dielectrics, metal gate electrodes, advanced implant and millisecond annealing just to name a few. As with the last technology node early engagement on these new materials and processes is critical and has already started and is beginning to accelerate. □