PVD REFLOW ENABLES VOID-FREE COPPER FILL

IN THIS ISSUE
• Cryo-Implantation in Ultra-Shallow Junction Formation
• High Aspect Ratio Etching for 3D NAND
• Developing ReRAM for Future Memory Applications
At the 2x nm node, the industry finds itself in the situation aptly described by baseball icon Yogi Berra: “when you come to a fork in the road, take it.” Transistor technology is advancing in multiple directions. While material properties and physical barriers are driving the growing trend from planar to 3D designs, much work is still being done to extend planar scaling and capability to the utmost.

Conductor etch is advancing on both fronts. Developments in high-temperature etching are extending 2D NAND Flash technology, while improvements in critical etch CD uniformity and microloading control address double- and quadruple patterning as well as high aspect ratio 3D architectures.

At the 2x nm node, void-free interconnect metallization becomes a major challenge as ionized physical vapor deposition (PVD) cannot guarantee the required coverage for electroplating. We present an innovative solution that enables bottom-up copper fill through a PVD reflow solution, effectively reducing via aspect ratios and facilitating the subsequent plating step.

Dopant activation and defect creation in the extension region, source/drain junction, and contact region can also impede performance scaling at 2x nm and beyond. Our leading-edge cryo-implantation process, taking place at wafer temperatures as low as -100˚C, combined with energy-pure dopant profiles is effectively addressing these challenges.

Emergent 3D NAND architectures heighten the requirements for high aspect ratio etching, demanding precise control of both plasma characteristics and reactor conditions. A magnetically tuned very-high-frequency source, combined with high-frequency bias and step-to-step temperature control enables all-in-one etching of both mask and dielectric stacks for very high aspect ratio etching and extremely high tungsten selectivity.

With continued planar scaling of the contact length, the interface resistance between metal and semiconductor has become a major challenge for CMOS performance and power scaling. We are studying metal-insulator-semiconductor contact architectures to determine the most effective way in which to enhance current-carrying capability to match state-of-the-art NiSi contact architecture.

Although 3D NAND appears to be the leading contender for next-generation bulk-storage memory, other architectures offer possible alternatives. Structural simplicity, cell scalability, switching speed, and compatibility with established CMOS fabrication processes make ReRAM an attractive complement to 3D NAND for future non-volatile memory applications at 1x nm and beyond.

We hope you find these articles thought-provoking springboards for discussion and collaboration in addressing the multiple challenges we face in advancing the state of the art in semiconductor technology.

Cover: Innovative copper seed technology for the 1x nm node combines cold PVD deposition and high-temperature copper reflow. An iterative cycle of the two steps reduces feature aspect ratio for electroplating, facilitating void-free fill.
PVD Reflow Enables Void-Free Copper Fill at the 2x nm Node and Beyond

Cu metallization in today’s microelectronic fabrication involves depositing the Cu barrier and seed layers by PVD followed by Cu electroplating to fill the topography. Each of these process steps requires careful optimization that ultimately determines the success or failure of the gap fill. Incremental enhancements of both barrier/seed and electroplating processes have ensured void-free Cu gap fill down to the 2x nm node. As interconnect structures shrink further, even the most optimized barrier/seed process, showing conformal coverage without any overhang, increases feature aspect ratios beyond levels manageable for electroplating (Figure 1). Reflow of sputtered Cu is one of the most promising approaches for resolving this issue in interconnect metallization.12 Via structures shrink further, even the most optimized barrier/seed process, showing conformal coverage without any overhang, increases feature aspect ratios beyond levels manageable for electroplating (Figure 1). By enabling bottom-up Cu fill and thereby reducing aspect ratio, it facilitates electroplating and void-free gap fill (Figure 1). Cu movement into narrow trenches and vias can be achieved by augmenting deposition with subsequent thermal reflow. Cu reflow is controlled by several fundamental processes. Surface diffusion is driven by capillary forces created by local gradients of chemical potential, which in turn vary due to surface curvature and geometry. Grain boundary movement accommodates material flux, from high potential to low potential, and from the convex side to the concave side. The wetting properties of the underlying layer also play a role. To avoid lateral growth and agglomeration of Cu after reflow, a selective deposition profile (thinner field coverage and thicker bottom coverage) is essential as it utilizes the above-mentioned phenomena to drive the Cu downwards into the narrow structures, creating a bottom-up PVD fill.

INNOVATIONS IN PVD CU REFLOW

Pedestal heaters are commonly used for wafer heating in a PVD deposition system. However, this approach poses significant process limitations for PVD reflow, which requires two process temperatures—cold deposition followed by high-temperature reflow. Instead, reflow can be induced through backside wafer heating using lamps, reflectors, and a wafer lift assembly embedded in the wafer pedestal. Figure 2 (a-c) details the process sequence of room temperature deposition and high-temperature reflow. The wafer enters the chamber cold and Cu deposition occurs at room temperature. Innovations in the PVD Cu deposition system (not discussed here) enable the selective Cu deposition profile essential for achieving bottom-up fill from the reflow step. Following deposition, the wafer is raised to the reflow position using an integrated pedestal lift assembly. The wafer reheats from the convex side to the concave side. The wetting material flux, from high potential to low potential, and from the convex side to the concave side, the wetting properties of the underlying layer also play a role. The above-mentioned temperature regimes also depend on the substrate layer for Cu reflow. Following reflow, the wafer is lowered onto the pedestal and electrostatically chucked to enable rapid cooling. The cycle of deposition, reflow, and cool-down can be repeated multiple times to gradually increase bottom coverage in features to the point of complete fill. Figure 2(b) shows complete Cu fill of 2x nm node trench structures achieved after three reflow cycles on Ta liners.

Figure 2. Process sequence of (a) Cu reflow showing (b) wafer position and lamp status and (c) temperature profile.

Figure 3. Reflow performance on 2x nm structures. (a) Effect of temperature on Cu reflow on Ta liners and (b) 100% bottom-up Cu fill on Ta liner after 3X cycles.
In addition to room temperature deposition followed by thermally-assisted reflow, chamber architecture accommodates deposition at higher temperatures (400-400°C), followed by thermally-assisted reflow. In the latter scenario, lamps can be used to heat the wafer prior to the deposition step. This technique makes it possible to apply reflow on other liner materials, such as cobalt and ruthenium, as shown in Figure 4.

Temperature uniformity across the wafer is essential for ensuring yield after reflow. Lamps (diameter and design) and reflectors (shape, location, and surface reflectivity) are key to optimizing temperature uniformity. Figure 5 illustrates the temperature profile of the wafer during each step of the reflow process, as measured by a 17pt thermocouple. This measurement was done for a room temperature deposition process with a reflow set point of 325°C. The temperature non-uniformity range, measured after heat up and soak, is less than 25°C. The figure also shows the temperature window for achieving full Cu reflow on a Ta liner (2x nm node trench structures).

DEVICE INTEGRATION AND MANUFACTURABILITY

The impact of a high-temperature reflow process on overall device performance was evaluated by line and via resistance measurements. No significant difference in resistance was found between test samples with and without reflow (Figure 6). Performance, however, is sensitive to the ion energy during the PVD Cu deposition process. Low-energy processing minimizes the intermixing of Cu and Ta during deposition. This reduces electron scattering along the Cu-Ta interface and maintains low line and via resistance.

High-volume manufacturability of the deposition and reflow process was validated in a 10,000-wafer marathon. Figure 7 shows stable thickness uniformity, repeatability, and an in-film particle performance of fewer than 10 adders at 0.09µm.

CONCLUSION

The technology node shrink from 2x nm to 1x nm creates significant gap fill challenges for PVD Cu barrier/seed and electroplating. Reflow-based PVD Cu seed is required to address Cu seed coverage challenges. A cycle of deposition, reflow, and cool-down performed in a single chamber demonstrated stable marathon performance in achieving complete gap fill on a PVD liner with extremely low defectivity. Line and via resistance measurements were comparable with baseline non-reflow PVD Cu samples. Independent temperature control for each step of the deposition/reflow cycle offered flexibility for successfully applying this process to cobalt and ruthenium liners.

ACKNOWLEDGEMENTS

The authors acknowledge contributions from the technology, engineering, and integration groups of the Applied Materials Metal Deposition Products team.

REFERENCES


AUTHORS

Sree Kesapragada is a global product manager in the Metal Deposition Products business unit of the Silicon Systems Group at Applied Materials. He holds his Ph.D. in materials science and engineering from Rensselaer Polytechnic Institute.

Winsor Lam is a technology manager in the Metal Deposition Products business unit of the Silicon Systems Group at Applied Materials. He earned his B.S. in chemical engineering from the University of California at Berkeley.

Xiaomin Tang is a principal member of technical staff in the Metal Deposition Products business unit of the Silicon Systems Group at Applied Materials. He received his Ph.D. in physics from the College of William and Mary.

Mani Subramani is a distinguished member of technical staff in the Metal Deposition Products business unit of the Silicon Systems Group at Applied Materials. He received his Ph.D. in mechanical engineering from the University of Rhode Island.

Keith Miller is a senior engineering manager in the Metal Deposition Products business unit of the Silicon Systems Group at Applied Materials. He holds his M.S. in mechanical engineering from Purdue University.

ARTICLE CONTACT

Sree_Kesapragada@amat.com

PROCESS SYSTEM USED IN STUDY

Applied Endura® Amber™ PVD
Cryo-Implantation in Ultra-Shallow Junction Formation at the 2x nm Node and Beyond

A major development in this field during early 2x technology development has been the addition of wafer temperature as a means of further improving junction formation. Cryo-implantation, which involves implantation at wafer temperatures as low as -100°C, is now being implemented in fabs of the most advanced device makers. Extensively studied, this method is employed in a number of device applications, showing significant benefits in reducing crystal damage, improving junction engineering, alleviating strain relaxation in SiGe (silicon germanium), and reducing device leakage.

REDDUCING DAMAGE

At the 2x nm node, junction depths must be less than 15nm to minimize short channel effects (SCE). Junction depth is driven primarily by the enhanced dopant diffusion caused by the formation and evolution of defects that advance into the silicon upon annealing. Substrate temperature exerts a powerful effect on the implantation process. Cooling the wafer during implant it has been demonstrated that fewer self interstitials (target interstitials) are present at the implant end-of-range (EOR), i.e., the depth just beyond the peak of the implanted dopant profile. Lower temperatures down to -100°C are typically more effective. At these temperatures, the lattice energy is reduced along with its ability to recover, thus accelerating the rate of lattice disruption caused by implant. Under these conditions, amorphization occurs at lower doses and extends deeper with fewer self interstitials left beyond the amorphous/crystalline (a/c) interface. The reduction of EOR defects formed upon annealing as post-implant thermal budget decreases. Phosphorus also diffuses more rapidly than arsenic, but the smaller contact areas at advanced nodes dictate higher active dopant abruptness and concentration to reduce series resistance. Cryo-implantation is an attractive alternative as its activation level and solubility are superior to those of arsenic, but its lighter mass results in more channeling. Pre-amorphization can suppress channeling, but inevitably leads to damage, which is less effectively annealed as post-implant thermal budget decreases.

Co-implanting the carbon at -100°C removes the need for the carbon to be effective. However, traditional pre-amorphizing implant can cause significant EOR damage and high leakage. Co-implanting the carbon at -100°C removes the need for pre-amorphization and enables very abrupt, low-resistance junctions using phosphorus, as illustrated in Figure 2. Using the implanted carbon plus phosphorus profiles in Figure 2a cryogenic carbon implantation achieved an 8nm/dec abrupt junction as shown in Figure 2b.

Figure 1. Cross-sectional TEM images for (a) boron 2keV 3e15/cm² at RT and -100°C, (b) carbon 5keV 3e15/cm² at RT, -20°C, -40°C, and -100°C reveal thicker amorphous layers and smoother a/c interfaces with cryo-implantation. The effect is more evident as temperature decreases.

Figure 2. Implant profiles of phosphorus, boron, and carbon after (a) cryo-implantation and (b) annealing.

At wafer temperatures as low as -100°C is proving a versatile technique for enabling continued device scaling. Scaling for smaller and faster devices, with the accompanying trends to reduce diffusion and lower thermal budgets, have driven advances in ion implantation technology. Ultra-shallow junction formation in the source/drain extension (SDE) requires extremely abrupt profiles, aided by the elimination of energy contamination in the implanted profiles as well as means to suppress diffusion. Incorporating strain to independently optimize performance of n-type and p-type devices is a complex process made easier and more effective by ion implant’s precise placement of material-modifying species. Contact area becomes so small that contact resistance severely inhibits device performance, challenging integration and process engineers to increase dopant activation without incurring major process changes.

A major development in this field during early 2x technology development has been the addition of wafer temperature as a means of further improving junction formation. Cryo-implantation, which involves implantation at wafer temperatures as low as -100°C, is now being implemented in fabs of the most advanced device makers. Extensively studied, this method is employed in a number of device applications, showing significant benefits in reducing crystal damage, improving junction engineering, alleviating strain relaxation in SiGe (silicon germanium), and reducing device leakage. Scaling for smaller and faster devices, with the accompanying trends to reduce diffusion and lower thermal budgets, have driven advances in ion implantation technology. Ultra-shallow junction formation in the source/drain extension (SDE) requires extremely abrupt profiles, aided by the elimination of energy contamination in the implanted profiles as well as means to suppress diffusion. Incorporating strain to independently optimize performance of n-type and p-type devices is a complex process made easier and more effective by ion implant’s precise placement of material-modifying species. Contact area becomes so small that contact resistance severely inhibits device performance, challenging integration and process engineers to increase dopant activation without incurring major process changes.

KEYWORDS
Cryo-Implantation
Ultra-Shallow Junction Formation
Strain Relaxation
Leakage
Carbon Implantation

The reduction in junction depth and gate pitch that has accompanied advanced device scaling has required balancing the need to increase the depth of the SDE to lower resistance and the need for a shallow SDE to improve device SCE performance. Conventional nMOS contacts have been made using arsenic ion implants, but smaller contact areas at advanced nodes dictate higher active dopant abruptness and concentration to reduce series resistance. Phosphorus is an attractive alternative as its activation level and solubility are superior to those of arsenic, but its lighter mass results in more channeling. Pre-amorphization can suppress channeling, but inevitably leads to damage, which is less effectively annealed as post-implant thermal budget decreases. Phosphorus also diffuses more rapidly than arsenic, producing poor junction abruptness and the inability to achieve ultra-shallow junctions for the sub-20nm nodes. Introducing carbon at the junction reduces this effect, but typically silicon must be fully amorphized for the carbon to be effective. However, traditional pre-amorphizing implant can cause significant EOR damage and high leakage.

Figure 1. Cross-sectional TEM images for (a) boron 2keV 3e15/cm² at RT and -100°C, (b) carbon 5keV 3e15/cm² at RT, -20°C, -40°C, and -100°C reveal thicker amorphous layers and smoother a/c interfaces with cryo-implantation. The effect is more evident as temperature decreases.

Figure 2. Implant profiles of phosphorus, boron, and carbon after (a) cryo-implantation and (b) annealing.
Cryo-Implantation

For comparison, Figure 2c shows the broader phosphorous profile and deeper junction formed without cryo-implantation. Moreover, the 80nm gate pitch nFETs (n-type field effect transistors) tested in this study displayed resistance values 15% lower than those doped with arsenic, inducing an 11% increase in current (I_n) for matched I_f (Figure 2d).

PRESERVING STRAIN, REDUCING LEAKAGE

At advanced nodes, channel strain engineering enhances carrier mobility in pMOS transistors. In these devices, compressive strain is induced by SiGe selectively grown in the source/drain regions.[2] For maximum effect, the SiGe must retain the same level of strain through multiple downstream process steps, including implantation and annealing.

It has been demonstrated that strained SiGe relaxes via misfit and threading dislocation propagation after high-temperature dopant-activation annealing. EDR defects, which commonly form around the a/c interface during annealing can serve as nucleation sites for dislocation formation or in the inter-diffusion of boron and germanium that, in turn, can accelerate strain relaxation.

Cryo-implantation reduces the number of EDR defects, which has the dual benefit of preserving strain and reducing the likelihood of junction leakage emanating from EDR defects. Results from one study (Figure 3a) showed that pre-silicidation strain decreased by approximately 7% in the absence of cryo-implantation, but was maintained when carbon cryo-implant was implemented.[2] In another (Figure 3b), 28nm CMOS structures in which p/n junctions were formed by in-situ doped epitaxial SiGe followed by germanium implants of varying energies revealed leakage two orders of magnitude lower using a 10keV germanium implant at cryogenic temperature.[2]

OPTIMIZING INTERFACES

Cryo-implantation has also shown good results in improving nickel silicidation interface qualities. Nickel silicide has been used with SiGe stressors since the 90nm technology node to reduce source/drain contact resistance. However, such issues as controlling nickel diffusion, avoiding nickel piping (usually along implant-induced damage sites) and controlling agglomeration, particularly in the SiGe strained layer, make silicidation especially challenging, particularly in p-type structures.[2]

As germanium concentration increases beyond 25% to enhance pFET channel stress, the nickel silicide/silicon interface tends to become rough. A pre-silicide cryo-implant mitigates this effect (Figure 4a). In addition, cryo-implantation enables pMOS depth and abruptness requirements by trapping interstitials and limiting nickel diffusivity. Figure 4b illustrates nickel silicidation formation at lower temperatures, with better thermal stability and lower sheet resistance when preceded by a -100°C carbon implant. For nMOS, nickel diffusion during the silicidation step can be reduced by using a silicon implant at cryo-temperature to pre-amorphize the contact area.

CONCLUSION

Ion implantation technology has evolved on several fronts in response to scaling challenges at the 2x nm node and beyond. Cryo-implantation in combination with energy-pure dopant profiles and special species, such as carbon and germanium, is proving to be a versatile and effective means by which to create ultra-shallow junctions with minimal damage, improved activation, and reduced leakage. With such advances, high-current ion implantation has become a true enabler of device performance and yield improvements for the 2x nm node and beyond.

REFERENCES

Critical trench etching processes, such as shallow trench isolation (STI) and combined STI/film formation for FinFETs now require profile and microloading control for features with aspect ratios (AR) exceeding 15:1 at the 20nm node and 20:1 at 15nm. These challenges are exacerbated by small variations in CDs of adjacent features created by the aforementioned advanced patterning schemes. Control of intra-cell microloading is a key issue and requirement in advanced IC manufacturing. Conventional ICP etch systems utilizing a single frequency RF bias (typically 13.56MHz) are limited in addressing these issues. Experiments have shown that dual frequency bias technology employing both 13.56MHz and 2MHz RF power to the cathode (wafer) offers flexibility in tuning the ion energy and angular distributions (IEADs) for tighter AR and microloading control. RF pulsing enhances this IEAD control through its ability to affect ARDE.[23]

High-κ/metal gate (HKMG) stacks were first introduced for logic gate applications at the 45nm node[24] and have been widely employed in logic designs from 32nm. NAND Flash scaling challenges led to the introduction of floating gate devices in which high-κ materials were integrated into the inter-poly dielectric (IPD) to maintain the coupling ratio between the control gate (CG) and floating gate (FG) while eliminating the CG wrap around. For this gate structure, in which the high-κ dielectric is significantly thicker than in logic gates, high-temperature etching appears to be the only way to maintain vertical profiles with minimal footing while stopping on extremely thin tunneling oxide without pitting in the substrate.

**ADVANCED PATTERNING**

The tolerable CD variation for feature sizes below 20nm is on the order of 1nm or less. Past development has focused on gas distribution (center feed with directional control) and plasma shaping to eliminate inherent “M-shape” plasma non-uniformities associated with conventional ICP sources.[25] With these improvements,
ICP-based conductor etch chambers, such as the one depicted in Figure 1, provide excellent control of ion and neutral species across the wafer. The focus has now shifted to the edge of the wafer, where discontinuities in material and geometry give rise to variations in electrical field that result in "roll-off" of CD and etch depth.

We have found that extending the electrical field out past the edge of the wafer mitigates such variations and results in tighter CD control. Figure 2 illustrates this improvement for an advanced patterning scheme over results obtained using a standard electrostatic chuck configuration.

**CRITICAL TRENCH ETCH**

Etching the silicon substrate for trench isolation has been implemented in numerous generations of IC manufacturing. As devices become more three-dimensional, trench etching now includes buried word line, buried bit line, combined STI/fin formation for FinFETs, and memory core STI, in addition to standard STI for transistor isolation. As we approach 20nm for memory planar NAND Flash stack as proposed by Parat et al., higher temperatures.

Unlike patterning and critical trench etch processes that are performed within the conventional temperature regime (25-80°C), satisfying advanced etch requirements for non-volatile materials used in HKMG requires much higher temperatures. Figure 5 illustrates an advanced planar NAND Flash stack as proposed by Parat et al. for sub-30nm half-pitch, which includes a high-κ IPD. Once the top layers have been etched, it is challenging to etch the bottom layers while preserving the top profile. Since the AR in the cell area can be equal to or greater than 10:1, the ion energy must be increased for conventional etch processes. Etch by-products of high-κ films (e.g., HfO2, La2O3, Al2O3) are non-volatile in the conventional operating temperature regime. Thus, etching high-κ materials at low temperatures requires a high sputter yield (high ion energy) physical etch process that typically results in poor selectivity to the underlying tunnel oxide and source/drain structures, as well as an undesirable tapered profile and footing (Figure 5). These adverse side-effects significantly constrain the process window.

To achieve high selectivity and vertical profiles in high-κ materials requires a high-temperature chemical etch regime. At temperatures of 150-250°C, little or no ion energy is required, resulting in near-infinite selectivity between high-κ and silicon oxide films. Figure 6 illustrates this effect, comparing HfO2 etch rates as a function of temperature for three ion energy (DC bias) levels. As expected, etch rates are higher at high DC bias, however, etch rate dependence on DC bias is noticeably reduced at higher temperatures.
wearer in the high-temperature regime compared with lower temperatures, creating a wider process window. The high selectivity process can be tuned to produce near-vertical profiles with minimal or no footing, which results in much lower word line to word line coupling. From a device reliability standpoint, this could be a critical requirement for better data retention and read/write program cycles.

CONCLUSION
Conductor etch applications face steep challenges as devices scale to 20nm and beyond. Critical dimension uniformity, microloading, and profile control requirements are becoming more stringent in advanced patterning schemes and three-dimensional architectures, as well as in next-generation planar structures incorporating high-k films. Etch technology development is showing good results in addressing these issues. Enhancing RF bias coupling to the extreme edge of the wafer can significantly reduce CD roll-off and etch depth variation that can result from discontinuity of the electric field. High AR trench etching benefits from the broader IEDF created by a dual frequency bias and from plasma pulsing. The former enhances etch depth, while the latter improves CD control and minimizes microloading. Planar NAND Flash devices present the dual etch challenge of higher AR features and relatively thick high-k material. As NAND Flash scales beyond 20nm, high-temperature etching is proving instrumental in obtaining the high selectivity and vertical profiles essential for optimal device functionality and associated device reliability.

REFERENCES

As the planar 2D NAND technology node shrinks, the cost of further advances (especially in lithography) becomes prohibitive and significant material-related challenges prevent scaling. This has expedited the need for 3D NAND. Current 3D NAND architectures involve multi-layer stack deposition, intensifying the demands placed on high aspect ratio (HAR) etching, including very HAR (>50:1) etch requirements, high throughput, and very high tungsten (W) selectivity for stanadce contact. Meeting these challenges, magneto chrome luminescence, very high frequency (VHF) plasma, and high material wafer thermal management deliver >7µm/min hard mask etch rates (ER), in-situ mask/lod etch with vertical profiles, and <500:1 W selectivity.

Conventional planar Flash memory technology is fast approaching critical and prohibitive scaling limitations. Some of these pertain to technical issues, such as integration density beyond the 45nm mode restricted by large cell size and high applied voltages/fields for program/erase operations. Others pertain to manufacturability, cost of scaling, and materials requirements. These fundamental limitations are inherently addressed by the transition to a 3D architecture and, hence, make 3D NAND more attractive. Stacking cells vertically creates a higher capacity/volume ratio in a smaller physical space and also improves electrical performance by shortening the interconnect between cells (which also reduces power consumption). While the progression of planar Flash is deterred by the cost of scaling lithography, 3D architecture can enable enhanced device performance using existing lower lithography/technology nodes in conjunction with deposition and etch requirements that are less expensive to realize.

For these reasons, 3D stackable NAND Flash memory is a mere two to three years from full-scale adoption. Most leading memory manufacturers expect to be in pilot production within the next year and in volume manufacturing within the next two years. Each manufacturer is investigating different architectures with different variations in the integration scheme: gate-first/gate-last, transistor architecture, and storage mechanisms. However, the common element is a stackable device structure of alternating layers of dielectric and conductor. The two key approaches are an oxide-poly alternating stack and an oxide-nitride alternating stack with wet etch removal of the nitride to fill with tungsten. Given the greater vertical dimension of the stack (AR=50:1) compared with planar Flash, both approaches place stringent demands on deposition and etch performance—in particular, on HAR etch.

3D NAND HAR ETCH CHALLENGES
As the aspect ratio of etch stacks increases, process tolerances become extremely tight. Key challenges include ER/aspect ratio dependent etching (ARDE) improvement, [2-3] critical dimension (CD) uniformity, [7] and CD profile improvements. Vertical profiles are especially critical (i.e., high bottom to top CD ratio, absence of bowing or bending, and distortion-free bottom holes). In addition, for the staircase contact architecture in 3D NAND, it is essential to avoid punch-through of the gate W during the entire etch. This implies that the W layer at the top of the staircase will experience more than 300% over-etching, effectively increasing the W selectivity requirement to >50:1. The advent of 3D NAND has also increased the aspect ratio of mask etching and overall mask etch requirements for ER, selectivity, and deformation.

KEYWORDS
High Aspect Ratio 3D NAND
VMR Plasma Etch
ARDE
Selectivity
Mask
To deliver the necessary etch performance to meet the above requirements, the plasma etch reactor must create a very uniform high-density plasma, higher ion energy, uniform neutral species distribution, and uniform wafer temperature. Typically, multi-frequency capacitively coupled plasma (CCP) sources have been used for HAR applications, using a higher radio frequency (RF) power to control ion density and radical fluxes, and lower frequency to control ion energy distribution.\(^\text{11}\) \(^\text{12}\)

**HAR ETCH INNOVATIONS**

A triple-frequency capacitive-coupled etch system has been developed that addresses the aggressive etch performance requirements for 3D NAND technology (Figure 1).

To generate uniform high-density plasma over the wafer area, both high-frequency (HF) RF and low-frequency (LF) RF are delivered to the cathode. To control dissociation of the process gases, a VHF RF source is configured to the top electrode. Multi-zone gas control and independent gas injection are used to achieve uniform and tunable neutral species distribution. The top source electrode and reactor walls are maintained at elevated temperatures to manage polymer distribution and deposition. Step-to-step active wafer temperature control (Figure 2) enables constant temperature within a process step—even under high plasma induced thermal loads—and step-to-step temperature tunability.

**HAR Etching of Multi-Stack Dielectrics**

Vacuum delivery and innovatively coupling different RF frequencies are crucial for achieving the ER, profile, and uniformity requirements in HAR etching. Applying HF bias power increases plasma density at the wafer level, while lowering the sheath voltage (Vrf) or plasma heating. Figure 2 shows the reduction in Vrf when using a higher frequency bias as compared to a 13.56MHz bias reference. Reducing the Vrf enables use of higher LF bias power—the primary means of increasing the ion bombardment energy—thereby promoting vertical profiles.

Supplementing the bias source with a VHF top source increases the plasma density further and helps achieve faster ERs. With VHF RF only, the plasma density is greater at the center of the chamber. As HF bias RF is added, the plasma spreads out toward the edge, creating more uniform density distribution. As the bias begins to dominate, the peak plasma density moves out to the wafer edge. This plasma density modulation through RF power application can be used as a means of tuning on-wafer ER uniformity (Figure 2b). ARDE is a fundamental challenge for very HAR etching and is more pronounced with the alternating layers of dielectric and conductor in the 3D NAND stack. Etch depth capability depends on the incident fluxes of fluorocarbon reactive species (including both the ions and the neutral radicals) and the surface reaction probability, or the net energy, at the bottom of the hole. Enhancing ARDE as required in 3D NAND staircase contact etch requires higher incident ion energy (driven by higher LF RF bias power), as well as higher ion flux and lower surface temperature to enhance the surface reactivity on the bottom of the hole. Maintaining a stable wafer surface temperature under high bias power conditions is very difficult as the high thermal load from the plasma can overwhelm conventional wafer cooling systems. Figure 3 shows the effect of not maintaining surface temperature under high bias power conditions. Temperature maintained for each process step; Figure 4 shows the schematic of the staircase contact.

**Transitioning to HAR Mask Etching**

As advanced patterning film (APF) mask thicknesses have grown to exceed 1μm, maintaining adequate mask selectivity and integrity are substantially more difficult yet vital for very HAR processes. Precise control of polymer deposition and elimination of extra-reactive species at the wafer surface can enhance mask integrity, reducing deformation and profile bowing while reducing selectivity at the bottom of the hole. Optimized gas flow during the process increases mask selectivity and limits necking polymer deposition on top of the mask, while scavenging species reduce fluorine radicals at the wafer surface.

Another challenge in 3D NAND staircase contact etch is W erosion. It is proportional to an opposing combination of surface reactivity to fluorine chemistries and the rate of polymer deposition (passivation). Hence, controlling polymer deposition is critical for achieving high W selectivity. Higher wafer temperature facilitates high selectivity by both decreasing the reactivity and increasing the polymer mobility to ensure sufficient coverage at the bottom of the hole. However, it is challenging to maintain HAR etch capability at a higher wafer temperature and polymer deposition rate; the latter can produce etch stop as a result of polymer clogging at the mask sidewall. Thus, preventing temperature creep with the plasma thermal load is critical for staircase etch management.

Conventional wafer cooling systems are severely limited in this respect and active wafer cooling is pivotal for meeting selectivity demands. Figure 3a details the temperature maintained for each process step; Figure 4 shows the schematic of the staircase contact.
Polymization can also be reduced by low-temperature etching. (This contrasts with the high temperature requirement for W selectivity in dielectric etching.) Active wafer temperature control enables consistent and stable low temperatures even at high source powers and plasma loads. Step-to-step control (Figure 3a) allows for quick changes in electrostatic (ESC) temperature from process to process, thus enabling in-situ, all-in-one plasma etch of mask and dielectric stacks.

APF etching also requires high plasma densities. These can be achieved by increasing the RF power to the system. However, substantially raising bias power for this purpose would lead to degradation of other mask attributes, such as profile and selectivity. Here, VHF tuning offers a distinct advantage, enabling high plasma densities that significantly raise the ER and throughput of the mask-open process (Figure 5) with no adverse side-effects.

Typically, VHF sources create center-high plasma densities and are prone to inherent density non-uniformities. As noted above, adding RF bias power can help offset this. However, for APF etching, high bias power is detrimental to profiles. The addition of uniformity tuning mechanisms, such as solenoid magnetic coils and a charged species tuning unit, help to alleviate this issue and achieve uniform CD and profiles during the mask-open process.

Controlling uniformity of sidewall passivation is another important component of uniformity tuning. Mask sidewall protection is addressed by tuning wafer temperature using a dual-zone ESC. Figure 6 shows the effect of tuning the center and edge zones on wafer temperature and mask CD bias uniformity.

CONCLUSION

Transferring to 3D NAND architecture creates significant challenges in dielectric etching, especially for HAR etching. Addressing them requires precise control of both plasma characteristics and reactor conditions. Innovative solutions that combine a magnetically tuned VHF source, HF bias, and step-to-step active temperature control are proving successful in doing so. The source enables high throughputs in mask etching, and high plasma density and polymerization for the dielectric etch process. HF bias improves plasma density and ion energy control to optimize ARDE. Besides facilitating precise polymer management to simultaneously achieve very HAR etching and very high W selectivity, step-to-step wafer temperature control enables in-situ all-in-one etching of both mask and dielectric stacks.

ACKNOWLEDGEMENTS

The authors acknowledge contributions from the technology, engineering, and product management groups of the Applied Materials Dielectric Etch team.

REFERENCES

BENCHMARKING RESULTS

The benchmarking studies reported here revealed that the lower specific contact resistivity observed in MIS contact structures as compared to metal-semiconductor (MIS) contacts can be attributed to modification of the MS work function difference. This modification can result from (1) dipole formation at the metal–insulator interface; (2) reduction of insulator/semiconductor interface charges due to attenuation of metal-induced gap states (MIGS)\(^\text{(17)}\); (3) reduction of insulator/semiconductor interface charges due to disorder-induced gap states (DIGS)\(^\text{(18)}\); and/or (4) dipole formation at the metal–insulator interface. This modification can result from (1) dipole formation at the metal–insulator interface; (2) reduction of insulator/semiconductor interface charges due to attenuation of metal-induced gap states (MIGS)\(^\text{(17)}\); (3) reduction of insulator/semiconductor interface charges due to disorder-induced gap states (DIGS)\(^\text{(18)}\); and/or (4) dipole formation at the insulator/semiconductor interface. This modification can result from (1) dipole formation at the metal–insulator interface; (2) reduction of insulator/semiconductor interface charges due to attenuation of metal-induced gap states (MIGS)\(^\text{(17)}\); (3) reduction of insulator/semiconductor interface charges due to disorder-induced gap states (DIGS)\(^\text{(18)}\); and/or (4) dipole formation at the insulator/semiconductor interface. This modification can result from (1) dipole formation at the metal–insulator interface; (2) reduction of insulator/semiconductor interface charges due to attenuation of metal-induced gap states (MIGS)\(^\text{(17)}\); (3) reduction of insulator/semiconductor interface charges due to disorder-induced gap states (DIGS)\(^\text{(18)}\); and/or (4) dipole formation at the insulator/semiconductor interface.
Using data from various sources, contacts to n-Ge were also benchmarked using the J vs. Np plot in Figure 3. \[1\] [19]\[22]\[24]

Despite the remarkable improvement achieved by using MIS architecture, as opposed to MS architecture, MIS contacts to n-Ge still lag the state-of-the-art NiSi/n-Si reference system.

These studies showed that inserting an insulator between the metal and n-Ge causes the insulator energy barrier to attenuate J. This is demonstrated by the TiO$_2$/n-Ge\[10\], AlO$_x$/n-Ge\[10\], and MgO/n-Ge\[10\] data points below the reference line in Figure 5. The most promising result was Al/TiO$_2$/n-Ge at a doping level of 10$^{14}$ cm$^{-2}$ (Figure 5).\[17\]

**CONCLUSION**

RC is one of the biggest challenges facing CMOS performance and power scaling. MIS contact architecture has been proposed for improving the RC to n-Si and n-Ge. To test the validity of this proposal, J vs. Np data from several studies of MIS contact architectures were benchmarked against the state-of-the-art NiSi/n-Si reference.

For silicon, TiN/Ta$_2$O$_5$/n-Si performed better than the reference. In the n-Ge case, the most promising result was Al/TiO$_2$/n-Ge, which might be possible to improve further by reducing DIGS.

**ACKNOWLEDGEMENTS**

The author thanks Jerry Gelatos, C-P Chang, and Balaji Chandrasekaran for technical discussion, and Satheesh Kuppurao, Sundar Ramamurthy, and David Kyser for managerial support.

**REFERENCES**


**AUTHOR**

Khaled Ahmed is a distinguished member of technical staff in the Silicon Systems Group at Applied Materials. He holds his Ph.D. in electrical engineering from North Carolina State University.

**ARTICLE CONTACT**

Khaled.Ahmed@amat.com
Developing ReRAM for Future Memory Applications

NAND (“not and” logic gate) Flash has become the leading representative of bulk-storage memory given the advantages of its non-volatility and low cost per bit. Although NAND speed is limited by fundamental operational characteristics, such as Fowler-Nordheim tunneling, its low-power operation has made it attractive for incorporation into both mobile phone and tablet products, two of the major semiconductor markets today. However, the speed with which NAND has scaled—one generation every 1.5 years—means that it will reach its geometric scaling limit (expected at 10nm) within the next 3 to 5 years.

In the research community, resistive random-access memory (ReRAM) was seen early on as a promising non-volatile memory technology that could succeed NAND. ReRAM sandwiches a resistive change material, such as a binary metal oxide (MeO), between two terminal electrodes. Its simple architecture, and the speed and power economy with which it operates make it ideal for information and communication technologies that have been evolving in the expanding world of mobile devices (Figure 1).[2-4]

Challenges persist in the rapid cost scaling that has been achieved with NAND technology, creating a considerable economic hurdle for ReRAM to clear. And once leading-edge device makers proposed a 3D version of NAND, in which the 2D NAND cell string was rotated 90˚ from horizontal to vertical, intense industry focus shifted away from ReRAM. Although 3D NAND solves some 2D NAND scaling issues, it still faces complex process challenges, such as creating high-mobility silicon channels, and the deposition and etch of extremely high-aspect-ratio (>50:1) features. If these are overcome in the near future, it will be more difficult for ReRAM to become cost-competitive with 3D NAND. However, if solutions prove more elusive, ReRAM may yet become the technology of choice in portions of the memory market.

As 2D NAND approaches its scaling limit, its successor is the subject of vigorous debate. Will it be 3D NAND or will other promising technologies establish a foothold? ReRAM has been overshadowed by 3D NAND, but its scalability, speed, low-power operation, endurance, and compatibility with established CMOS fabrication processes could make it viable in parts of the memory market of 16nm and beyond.

The negative resistance phenomenon had not received the attention it deserved until the early 1990s when it was observed that the electrical resistance of certain manganese compounds varied as a function of a current passed through them. The negative resistance phenomenon (also referred to as hysteresis) was observed in the 1960s in some material systems. These systems demonstrate reduced resistance when a current is applied. If, in their original state, they behave as dielectrics or insulators, the process of forming and breaking down the insulator and current to recover it.

In the research community, resistive random-access memory (ReRAM) was seen early on as a promising non-volatile memory technology that could succeed NAND. ReRAM sandwiches a resistive change material, such as a binary metal oxide (MeO), between two terminal electrodes. Its simple architecture, and the speed and power economy with which it operates make it ideal for information and communication technologies that have been evolving in the expanding world of mobile devices (Figure 1).[2-4]
ReRAM OPERATING MECHANISM

It is now widely accepted that the mechanism for MeOx ReRAM is filament based, as shown in Figure 3, in which a dielectric layer (orange color) is sandwiched in between a top electrode (TE) and bottom electrode (BE). By means of the so-called forming process, a conduction path in the dielectric layer is first created by breakdown, which can be achieved in many ways, such as time-dependent dielectric breakdown or field-induced breakdown.

Forming occurs only once to initiate the change in material resistance from a high-resistance state (HRS) to a low-resistance state (LRS). When a voltage is applied into the conductive path, a current is generated to circuit parasitic capacitance. A peak current created by capacitive discharge can greatly exceed the input; this exerts tremendous force on the filament, which is identified as the best candidate for the MeOx layer. Forming a HfOx-based memory.[8] Expertise developed in high-κ materials for CMOS gate and DRAM fabrication is very useful in designing ReRAM structures. In general, ReRAM’s fast response speed derives from oxygen vacancy diffusivity near the filament; hence, ionic motion of the material is a key consideration. At this point, though, no one material has been definitively identified as the best candidate for the MeOx layer.

Equally numerous are the number of possible electrode materials (e.g., TiN, TaN, Si for the bottom; Hf/TiN, Ti/TiN for the top), each of which raises specific interface sensitivities that must be thoroughly analyzed. For example, precise control of the MeOx interface will be crucial to avoid unintentional oxygen diffusion from the oxide memory element to the metal electrode, creating an unwanted insulating layer. And different metals will produce different film orientation, grain dimension, and interface roughness that will affect filament formation. [25] Further, the work function of the metal electrode also needs to be considered as the On/Off resistance can be modulated by a barrier created between the dielectric and electrode, and, more importantly, between the filament and electrode.

As a practical matter for bringing the technology to production, several aspects of ReRAM development must be greatly refined. Identifying the best materials systems from fundamental considerations is highly desirable; to date, reproducibility of results has been an issue as deposition, integration, and cell operation techniques vary. Stability of these systems with respect to memory endurance is another critical factor in selection. And, of course, cost will be a major determinant in filament choices. Moreover, as shown in Figure 2, ReRAM is being investigated in the context of multiple architectures that will ultimately have to converge.

Array implementation raises further issues. For example, current technologies have not yet realized fast, low-power, and stable operation simultaneously in these devices. Also, unipolar switching is needed if ReRAM is to be integrated in crosspoint array structures with a selector diode for high-density mass storage.[26] Switching is called unipolar when the set/reset switching operations can be obtained using the same voltage polarity, and bipolar when set and reset operations need opposite voltage polarity. However, unipolar operation requires higher current than bipolar operation. It is also less stable; reset energy can actually perform the set during operation.

In addition, for low-power operation, the forming voltage must be low. Fortunately, this can be relatively easily achieved in various oxide thin films by reducing their thickness. Carrying this notion to its conclusion, forming voltage can ultimately become comparable to the set voltage, eliminating the forming process.[27] The resultant lower switching power will translate into faster switching speed.

Thermal management inside and outside the ReRAM device will be a major challenge in developing high-density 1D/R (one diode, one resistor) ReRAM arrays. Investigating thermal properties in a heterostructure of metal and (semiconducting) insulator layers will push today’s leading-edge metrology to expand into 2D and 3D measurement.
CELL OPTIMIZATION
Optimizing the ReRAM cell will require attention to achieving low forming and reset voltages, low programming current, high endurance, and a large On and Off window for the resistance state. Scaling will lead to a phase-out of the forming process because the voltage decreases linearly with decreasing oxide thickness, while maintaining the same programming current.

Several studies have also demonstrated successful use of low-resistance oxides (e.g., oxygen deficient TiO₂) to effect a forming-free process. In the resist process, ionic motion of oxygen vacancy dominates in electrochemical removal of the conducting path, but accompanying tunneling and ohmic currents consume most of the reset power. To address this, the device structure must include a high work function metal to suppress the tunneling current.

Optimizing endurance will require tight control of oxygen movement at the interface between the electrode and the conducting oxides. One method of achieving this is to add a second, easily oxidized metal layer at the interface to serve as an oxygen reservoir and prevent oxygen penetrating into the electrode during the resistance switching. The On and Off window of the resistance state in ReRAM is larger than in other emerging non-volatile memories (e.g., spin transfer torque). Both this and the capacitance of the MIM stack will require refinement.

APPLICABLE PROCESS TECHNOLOGIES
From the film stack processing standpoint, avoiding oxidation of the metal during thin film growth is crucial. Therefore, tightly controlled chemistries and process conditions will be critical. A less reactive oxidant source, such as water, should be considered for binary oxide deposition. Uniform growth of films 2 to 10nm thick is typically required, for which current atomic layer deposition (ALD) processes are well suited. In addition, today’s chemical vapor deposition (CVD) and radio-frequency physical vapor deposition (RFPVD) processes can be readily tuned to control the composition, crystalline structure, surface morphology, and stoichiometry of the MeO₂ insulating films and metal nitride electrodes that, in turn, determine ReRAM performance and endurance. They operate in appropriately low pressure regimes for minimizing film damage and offer deposition rate control that enhances uniformity, while in-situ cleaning processes facilitate the required oxidation-free surfaces.

While each layer in the MIM stack is critical, other aspects of integration are equally important. For example, if the MIM cell is to be defined by a reactive ion etching process, the quality of the cell sidewall will vitally affect performance. This is especially important in the 1T1R structure (center image of Figure 21), in which the device size is on the order of 10 to 20nm and the aspect ratio of the structure is extremely high. Specially designed passivation will also be needed to reduce the charge or interface state that can affect memory properties, such as data retention.

Chemistry and plasma power will have to be carefully controlled to minimize damage of the stack during etch, and post-etch cleaning will also be essential. The chemistry of the latter will have to be constituted to avoid unintended chemical reactions with the cell. And, as aspect ratios become more aggressive, new cleaning techniques, such as gas chemical vapor cleaning, and thermally-assisted surface chemical reaction and volatiles, will be required to preserve the structural integrity of minute features at the 1nm node and beyond. Finally, a dielectric layer with low charge trap states will have to be applied to passivate the sidewalls. CVD and ALD will be key in fulfilling deposition requirements at these ultra-small geometries. As the integration scheme transitions to a more cost-effective structure (such as that illustrated on the right in Figure 21), processing requirements will become topographically sensitive. Both CVD and ALD will be used initially in achieving the requisite uniform film deposition, but as features become more challenging and uniformity requirements more stringent, ALD will be the only capable method.

Cluster processing will be essential for creating film stacks comprising tens of pairs of metal and dielectric layers with optimized interfaces and high process throughput. And etching 40nm contact openings at aspect ratios exceeding 50:1 through stacks of both metal and dielectric films will be exceptionally challenging, as the different processes will have to be carried out in one system using a range of chemistries and sources powers.

CONCLUSION
Structural simplicity, cell scalability, and switching speed make ReRAM an attractive complement to 3D NAND for future non-volatile memory applications. However, much materials and interface characterization remains to be done to arrive at an optimized materials system for the ReRAM cell and the most effective array structure for achieving the desired low-power operation for mobile devices. Fortunately, once these issues have been resolved, ReRAM fabrication will be feasible using established CMOS technologies, such as ALD, CVD, and RFPVD.

REFERENCES