Progress in nanoscale dry processes for fabrication of high-aspect-ratio features: How can we control critical dimension uniformity at the bottom?

Kenji Ishikawa1, Kazuhiro Karahashi2, Tatsuo Ishijima3, Sung Il Cho4, Simon Elliott5, Dennis Hausmann6, Dan Mocuta7, Aaron Wilson8, and Keizo Kinoshita2

1Nagoya University, Nagoya 464-8601, Japan
2Osaka University, Suita, Osaka 565-0871, Japan
3Kanazawa University, Kanazawa 920-1192, Japan
4Samsung Electronics Co., Ltd., Hwasung, Gyeonggi 445-701, Korea
5Tyndall National Institute, University College Cork, Cork T12 R5CP, Ireland
6Lam Research Corporation, Tualatin, OR 97062, U.S.A.
7Imec, B-3001 Leuven, Belgium
8Micron Technology Inc., Boise, ID 83707-0006, U.S.A.

*E-mail: ishikawa.kenji@nagoya-u.jp

Received March 16, 2018; accepted March 24, 2018; published online May 25, 2018

1. Introduction

Currently, feature miniaturization in semiconductor fabrication is continuing through newly proposed nanoscale and atomic layer control processes. In 1959, Richard Feynman made the now famous statement that “there is plenty of room at the bottom”, referring to the manipulation of matter on the atomic scale. In the same year, the first integrated circuit (IC) comprising germanium mesa bipolar transistors was developed. Shortly thereafter, the monolithic planar process was realized. In the 58 years since then, semiconductor technology has made astonishing progress, leading to today’s microprocessing units (MPUs) that comprise multiple billions of transistors with feature sizes quickly approaching atomic dimensions, as well as dynamic random access memory (DRAM) with 50 trillion one-transistor/one-capacitor (IT1C) structures fabricated in a diameter of 300 mm wafer. DRAM is a volatile memory, and the NOT-AND (NAND) type flash architecture using floating-gate transistors is an electronic nonvolatile memory. In 2008, an epoch-making three-dimensional (3D) gate stack NAND flash memory (3D-NAND) was developed to replace the planar two-dimensional (2D) flash memory. Today, flash memory cells with an impressive 512 GB per die are achieved by forming a 64-layer gate stack. The semiconductor memory market was expected to increase to close to $85 billion in 2017 and continue to grow to over $100 billion in 2020. Manufacturers are continuing to shrink feature sizes and to scale vertically to increase the bit density and meet demand while maintaining cost advantage over their competition. This is driving new challenges in dry etching processes to meet the demands of the ever-shrinking size and/or dramatically increasing aspect ratio of the features required for the next generation of chips. Fabrication of vertical gate stack structures requires HAR etching, resulting in dimension scaling from two to three dimensions (Fig. 1).

Developments in DRAM and 3D-NAND technologies have continued to scale through 3D integration and advanced pitch multiplication schemes. The capabilities of dry etching are continuing to downsize IC features from one generation to the next. The ability to make small features and control...
feature-size variation enables the fabrication of HAR features in high-volume manufacturing.

In complementary metal oxide semiconductor (CMOS) logic gates, since the advent of the production of fin field electric transistor (finFET) devices in 2011, device architectures have fundamentally adopted a 3D design.10 Process technology for etched vertical profiles with high aspect ratios (HARs), clean gate sidewalls free of etching process residue, minimal erosion of liner oxide films protecting key architectural elements such as fins, and residue-free corners at gate interfaces with critical device elements requires the establishment of design technology co-optimization (DTCO) or design technology for manufacturing (DFM) in device architectures and CMOS layout design for the perturbation-free process variations (Fig. 1)11.

First, we introduce the history of feature scaling, involving the transition from lithography scaling to film/etching-based multiple-patterning scaling, leading to the scaling from two to three dimensions. The critical dimension (CD), that is, the smallest geometrical feature, will be reduced in two dimensions by the introduction of higher-resolution lithographic tools with an extreme ultraviolet (EUV) light source, directed-self-assembly (DSA) materials, and other techniques. Second, we review the background and science of HAR etching technology, especially to give an atomic scale understanding of surface reactions at hole bottoms. The transport of charged particles (electrons; positive and negative ions) and neutral particles (“neutrals”) is reviewed to clarify the actual surface reactions. Third, macroscopic wafer-scale CD uniformity is examined to realize high-volume manufacturing. Traditionally, microloading effects, reactive ion etching (RIE) lag, and aspect-ratio-dependent etching (ARDE) have been studied. Under the HAR requirement, etched profile distortions become a significant issue for CD control at the bottom of features such as holes. Fourth, focus is placed on emerging processing technology and DFM. To control the CDs at the bottom of deep features, self-limiting reactions have attracted much attention.

Here we review nanoscale and atomic layer processing while focusing on the following topics: (1) advances in the development of atomic layer processing for HAR features achieving 2D to 3D scaling, (2) future challenges to controlling CDs, (3) CD uniformity at the feature and wafer scales, and (4) CDs at the bottom of deep features. The finFET is one of the leading 3D devices, but its fabrication faces the problems of uniformity in film deposition and material etching. In particular, using atomic layer deposition (ALD) and atomic layer etching (ALE) etching as boosting technologies, we discuss how the self-limiting surface reaction achieves CD control. Furthermore, we present an overview of possible technology ideas that can enable a continued increase in the number of functions per unit cost for the next two decades. The topic of DFM will be reviewed. One purpose of scaling is to realize new device features with low power consumption and high performance in terms of power, performance, area, cost, reliability, and manufacturability (PPACR). Accordingly, we should continue to explore new methods that provide solutions. The coordination of developments among technology, design, and systems, such as design technology system co-optimization (DTSCO), DTCO, and system and technology co-optimization (STCO), will enable us to overcome scaling hurdles more easily than by traditional methods. We also discuss new approaches to feature miniaturization.

2. Background on film/etching-based multiple-patterning scaling

The pattern density of memory devices has been dramatically increasing for decades. To integrate high-density circuits in a small chip, it was necessary to decrease feature sizes and increase the aspect ratios of interconnecting structures. Memory technologies are continuing to scale in dimensions. A 300-mm wafer now contains close to 400 die and 50 trillion features.6

In the past, the photolithographic CDs determined the feature dimensions. Since 2017, film and etching multiplicity technology has provided planar features with CDs smaller than the photolithographic resolution. DRAM is still scaling in two dimensions and the memory industry has adopted various pitch multiplication schemes to overcome the limitations of standard lithography.

Methods for producing sub-resolution mask patterns are roughly classified into (1) double exposure, (2) double patterning or the litho–etch–litho–etch (LELE) process, and (3) self-aligned spacer double patterning (Fig. 2). The main technique has become the spacer (sidewall) image transfer (SIT), as developed in 1984.12 The SIT process scheme is as follows. First, dummy mask patterns, for instance, developed resist-patterned masks, are coated by conformal oxide films. Then, the oxide films are etched off by anisotropic directional etching until the underlying substrate is exposed. The dummy masks leave surrounding collars of the oxides. The dummy masks are then removed so that only the collar features remain. Thus, the collar widths are controlled by the verticality of the sidewalls of the dummy mask and the conformality of the sidewall films. The pattern density can be doubled by the SIT method.13–16 The SIT patterning process is well adapted to state-of-the-art nanoelectronics developments.17–21 Furthermore, the double patterning can be conducted twice to achieve quadruple
Current lithography technology has enabled the patterning of high-density sub-15-nm features for DRAM. Although the SIT method achieves smaller CDs, its drawback is the process complexity of the film/etching-based multiple patterning. To reduce the process complexity and cost, the adoption of EUV in memory manufacturing lithography, motivated by cost reduction in high-volume manufacturing, has been postponed until around 2021 (Fig. 3). Because the SIT process was adopted with a finer photolithographic pattern, the original EUV pattern dimensions will reduce the complexity of quadruple- and multiple-patterning processes. EUV photolithography is described elsewhere.

As discussed above, traditional scaling was driven by lithographic improvements to reduce the CD. In this first generation, photolithographic scaling was successful and improved device performances. Over the last decade, a second generation of multiple-patterning methods, such as the SIT process, has provided alternative ways of sub-resolution patterning for photolithography. In the next decade, EUV with double patterning will play a role and eventually replace the current multiple-patterning methods. In addition, the transition from two dimensions to three dimensions is continuing to decrease device scales, in combination with dry processes involving film deposition and material etching.

When EUV high-resolution patterning is achieved to meet HAR profile control, the limiting factors for downsizing future DRAM nodes will be the variations in the pitch multiplication process (Table I). Physical parameter variations such as CD result to electrical parameter variations such as capacitance, current conductance, and leakage. Thus, the manufacturing processes are affected the variations to the device performances such as power consumption, delay, and reliability. To determine correlations in the physical parameters with electrical parameters are made with modeling and it is used statistical process control. Most commonly, the variations are distributed with Gaussian distributions. As noted, the statistics are categorized in systematic and nonsystematic components. Even if the physical device parameters are normally distributed, the dependence of the electrical device parameters on these physical parameters may not be linear, giving rise to non-Gaussian.

Even in 2D scaling, CD uniformity is a key issue for 300 mm wafer production lines. The CD uniformity should be controlled at both the feature and wafer scales, where these variations are known as microloading effects and loading effects, respectively.

### 3. 2D CD control

#### 3.1 Wafer-scale control of planar CD uniformity

Block copolymers have received much attention as an approach to solving the CD uniformity problem. Bates et al. reported a thermodynamic mechanism for controlling the structure of block copolymerization. Kim et al. applied self-assembly to form block copolymers into a nanoscale mask with a cylindrical shape. Bang et al. introduced block copolymer nanolithography in 2007. Thereafter, Bencher et al. reported their trials for sub-lithographic patterning using a block copolymer mask. Servin et al. introduced a directed self-assembly (DSA) process to increase CD uniformity due to diblock copolymerization with a natural period. The nature of self-assembly is useful for the realization of a tremendously narrow CD distribution.

| Table I. Sources of variation in pitch multiplication process. |
|---------------------------------------------------------------|
| 1. Feature-to-feature CD variation or local CD line variation (LCV) |
| 2. Field-to-field within wafer CD (WIW) variation or across chip line variation (ACLV) |
| 3. Wafer-to-wafer CD (WTW) variation or across wafer line variation (AWLV) |
| 4. Lot-to-lot CD (LTL) variation or across lot line variation (ALLV) |
| 5. Chamber or scanner CD variation |

In the late 1990s, reactive ion etching (RIE) equipment, such as inductively coupled plasma (ICP) and electron cyclotron resonance (ECR) equipment, was introduced to provide a high-density plasma source. Soon afterward, charge build-up at the bottom of HAR mask patterns became an issue. This phenomenon is caused by the electron shading effect and it occurs when the charge neutralization at the bottom is unbalanced. In addition, this shading effect governs the charging damage of transistors, by conducting current from the plasma into the dielectrics.

Vertical isotropy of etched features is obtained by ion transport during plasma sheath formation. In principle, positive and negative particles should have the same trajectories inside a hole and equalize the charge at the HAR hole bottom. Alternatively, if negative ions can move inside an HAR hole and reach the bottom in the same way as positive ions, the charges will be well balanced and solve the charge build-up issue. However, efforts are continuing to achieve this balance.

In pulsed plasma excitation, during “off periods”, electron-negative molecules receive abundantly electrons to generate negative ions. As a result, the volume fills with positive and negative ions. This means that sheaths in front of the surface collapse and reduce the plasma potential. This behavior helps to solve the charge neutralization problem inside features. Pulse plasma technologies are reviewed elsewhere.

The transport of ions and neutrals inside HAR features is becoming important, and transport can be modified by...
reflection and adsorption of the species on the sidewalls of HAR features. Recently, the control settings of plasma equipment to achieve optimal charge transport have been debated, with some arguing for the synchronous application of bias power or a direct current (dc) bias.\textsuperscript{37–39} For instance, Tokashiki et al. reported that the adoption of synchronous pulse plasma etching technology improved the CD uniformity for large-diameter wafers owing to control of the ion flux and energy as well as radical-to-ion flux ratios.\textsuperscript{40}

In addition to charged particle transport, the wafer temperature plays an important role in the control of etched feature profiles and CD uniformity. Kanno et al. pointed out that the adsorption of etchants and inhibitors is influenced by temperature, and that a gate CD could be controlled by the wafer temperature.\textsuperscript{41} Milenin et al. reported that the electrostatic chuck temperature affected the CD uniformity.\textsuperscript{42} A multiple temperature-zone chuck enabled the control of CD drift and shift.\textsuperscript{42} Uniformity across the wafer scale is dominated by the spatial distribution of the plasma density. Maeda et al. reported that the plasma density uniformity was controlled by the magnetic field in an ECR plasma etcher.\textsuperscript{43} Kim et al. reported that the etching mask materials affected the polymer deposition performance; for instance, an amorphous carbon mask improved deposition as compared with amorphous silicon masks.\textsuperscript{44} Matsui et al. reported that dissociation of large-mass fluorocarbon molecules was suppressed by low-pressure pulsed plasma, leading to high-selectivity etching due to control of polymer deposition.\textsuperscript{45–47}

For the purpose of increasing wafer-to-wafer CD uniformity, the model prediction method was used to resolve the process drift or shift of CD uniformity.\textsuperscript{48} To date, various advanced process control (APC) approaches have been used to achieve CD uniformity in high-volume manufacturing.

Thus, uniformity within features and across a wafer has reportedly been controlled with respect to ions, neutrals, and reaction surface temperatures. In addition to surface reactions at the HAR bottom, fluxes and energies of charged particles, that is, electrons and positively and negatively charged ions, should be balanced to create a more neutral reaction environment. For ideal RIE, neutrals such as reactive radicals, “absorbrates”, play an important role in the synergism of the adsorbate surface under ion irradiation. Excessive deposition may distort the etched profile owing to by-products and redeposition, that is, “evacuation or exhaustion”, which is becoming more important for CD control (Fig. 4).

Presently, 3D scaling plays a key role in increasing device density. Nevertheless, the 2D downsizing of DRAM cells also increases the DRAM capacitor aspect ratio. At the same time, more stringent requirements for profile control and CD uniformity must be addressed. Here, focus is placed on (1) distortion of the etched profile and (2) CD uniformity at the HAR bottom. Etching rates have become so low that they are no longer economically viable. It is necessary to develop high-etching-rate methods without distortion of the etched profile. Overall, macroscopic etching rates depend on the aspect ratio and pattern, that is, the ARDE phenomenon. To compensate for ARDE, equipment development has focused on increasing the radio-frequency (RF) bias power and flow conductance.

Profile distortion effects continue to be seen in RIE lag, including notching, bowing, faceting, microtrenching, and profile shape dependence (Fig. 1).\textsuperscript{49} The charge build-up problem of the underlying layer during HAR etching causes etched profile distortions, as reported by many researchers.\textsuperscript{50–63} In the past, essential discussions have focused on electron shading effects and the trajectories of positively ionic species inside HAR features.\textsuperscript{58–64}

Sawin and co-workers argued for sidewall passivation in a Cl\textsubscript{2} plasma for Si etching in 1994.\textsuperscript{65,66} Haverlag et al. and Chang et al. reported that ion flux arrived both directly and indirectly at the feature bottom.\textsuperscript{67,68} The ion transport inside features predominantly determines the etched profiles as well as the profile distortions. Vitale et al. pointed out that the angular distribution of ion incidence on the bottom surface was affected by sidewall reflection of the ions.\textsuperscript{69} Mahorowala and Sawin reported the simulation of etched profiles in a mixed Cl\textsubscript{2} and HBr plasma for Si etching in 2002.\textsuperscript{70} Tsuda and co-workers provided computational simulation results for etched profiles.\textsuperscript{71,72}

The issue of ion incident angles has been investigated experimentally by various techniques. Using an ion beam etching system, Okano et al. reported Si etching rates and etching yield in a mixed Ar and Cl plasma.\textsuperscript{73} Using a Faraday cage, the Moon group researched the angular dependence of ion incidence on actual sidewalls.\textsuperscript{74–83} Recently, angled features were fabricated by a ClF\textsubscript{3}–Ar gas cluster injection system.\textsuperscript{84,85} Also, transport and redeposition of neutrals and by-products should be taken into consideration as causes of distortion. Nishikawa et al. reported in situ monitoring of etchant and product species in plasma etching using Fourier transform infrared spectroscopy (FT-IR).\textsuperscript{86} Morioka et al. argued that the desorption of by-products such as H\textsubscript{2} and SiCl\textsubscript{4} was limited to the reaction of Si in HCl and H\textsubscript{2} etching.\textsuperscript{87} Duluard et al. reported the formation of a passivation layer of Si-containing chloro-oxides at the sidewall of an etched feature at a low
temperature of $-20^\circ\mathrm{C}$.\textsuperscript{88} When by-product deposition onto the sidewalls occurred with aspect-ratio-dependent behavior,\textsuperscript{89,90} high bias power and pulse-modulated RF power reduced the dependence because the positive ions could not reach the bottom of the trench due to charging in the experimental results of a HBr and O$_2$ mixture plasma reported by Park et al.\textsuperscript{91} Iwase et al. recently reported results for HBr/N$_2$/fluorocarbon-based gas plasma and the effect of the wafer temperature on etching rates.\textsuperscript{92,93} We note the recurring themes of charging, absorbates, and redeposition of byproducts as being particularly important.

Enomoto et al. reported profile distortion in SiO$_2$ etching with CF$_4$ plasma in 1979.\textsuperscript{94} Oehrlein and Kurogi reviewed the SiO$_2$ sidewall chemistry in 1998.\textsuperscript{95} Ikegami et al. reported that bowing at the sidewalls occurred by reducing the ion flux and energy in HAR holes and by the bending of trajectories of the incident ions at mask edges.\textsuperscript{96} Izawa et al. found that radical transport inside HAR features affected the adhesion of etching protection films at the sidewalls due to less protection against ion-sputtering erosion.\textsuperscript{97} Miyake et al. pointed out that taper angle masks scattered incident ions, causing the bowing of sidewalls in HAR holes.\textsuperscript{98} Similar ion scattering at sidewall surfaces caused necking and bowing of the feature profiles, as found by the Moon group.\textsuperscript{92,93} Kamibayashi et al. discussed the effect of wafer bias frequencies on etched profiles.\textsuperscript{99} Negishi and co-workers reported that the mask edge roughness was transferred with amplification of the roughness amplitude, leading to deformed pattern profiles with profile distortion at the hole bottom.\textsuperscript{100,101} Such profile distortion problems have always been an issue for the plasma-etching fabrication of HAR features.

The historical trend of HAR contact hole (HARC) etching has been to develop equipment with higher bias power and larger flow conductance to etch deeper holes. As the aspect ratio of the hole increases, the etching rate dramatically decreases because the flux of ions and neutrals is too limited to reach the etching front.\textsuperscript{102} To conduct the highly selective etching required for SiO$_2$ over a thin masking photoresist film and the underlying films such as the Si substrate, W film, and Si$_3$N$_4$ film, dominant etching of SiO$_2$ is achieved by supplying radicals such as CF$_3$ species, and the reaction needs a relatively high energy induced by ion irradiation.\textsuperscript{102} This synergistic reactions of radicals and ions also requires sufficient amounts of reactive etchants (radicals and ions).\textsuperscript{103–106} Tatsumi et al. reported that etching rates under irradiation of highly energetic ions were determined by a number of F atoms in a few nanometer of a SiO$_2$ film to a covered polymeric layer which was formed by radical adsorption.\textsuperscript{107} This means that ion energy in sheath and adsorbed polymer layer thickness on the surface are correlated with etching rates, as a summary of several researches.\textsuperscript{108–111} In this situation, the wafer temperature has been the key cause of anomalous etched profiles with, for example, bowing and necking. At low wafer temperatures, excessive deposition occurs at the entrance of HAR holes and insufficient deposition occurs at the hole bottom. In contrast, at high wafer temperatures, the reduced sticking of CF$_3$ radicals increases the transport of radicals into deeper parts of the hole.\textsuperscript{116} The problem of absolute radical fluxes inside HAR holes has not been completely solved yet. Accordingly, HAR feature dimension control extends to the 3D-CD at the HAR bottom.

3.3 Radical transport and ion transport inside HAR features

The etching process for HARCs became a key process for memory devices, as described in Fig. 4. As memory technology moved on from laterally shrinking devices to vertically stacked devices, 3D-NAND was scaled up by adding more vertically integrated tiers of the oxide–poly–oxide–poly (OPOP) or oxide–nitride–oxide–nitride (ONON) films that are used to fabricate NAND memory cells.\textsuperscript{120} Therefore, the process capability for pillar etching is the limiting factor for increasing the bit density and reducing the per-bit node cost. As pillars are vertically scaled and the aspect ratio increases, the demands on the hard-mask technology and profile control increase and require the development of new processing technologies to maintain the CDs of the NAND memory cell throughout the pillar height.

Ion energy control by increasing the effective bias power for HAR features continues to advance. Great effort was made to increase ion energies to overcome charge build-up on the etching front inside HAR holes. Based on the trend of bias power in the past few years, the required power this year will exceed 25 kW, as shown in Fig. 5.\textsuperscript{121}

The ion energy distribution function (IED) depends on the frequency of RF excitation power with the bias power remains the same. For simplification, the ion energies are considered here on the basis of the effective bias power required to produce the same ion energy at a frequency of 2 MHz, regardless of the actual RF excitation power. To control the ion energy at an HAR hole bottom, higher bias power is becoming a critical technology for dielectric etchers.\textsuperscript{121} A new boosting technology is needed.

Many challenges arise when the bias power is increased. The prevention of arcing and effective cooling, and power delivery systems are all critical for enabling high-power capability. Moreover, it is harder to compensate for neutral fluxes with increasing aspect ratio, because the neutral species are transferred only by diffusion through the holes.

Pulse bias technology, which simply ramps up the bias power, was introduced almost 10 years ago.\textsuperscript{36,122–124} In the

![Fig. 5. Trend of bias power in reactive ion dielectric etchers during the past 10 years. Data are from two vendors of plasma etchers are shown.\textsuperscript{121}](image-url)
off period of the pulse operation, the sheath collapses and the potential at the wafer surface decreases, which allows electrons and negative ions to reach the etching front, thus neutralizing the positive charge build-up. This charge neutralization helps to increase the effective bias power and suppresses the accumulation of positive charges at the hole bottom owing to the transport of negative ions during the pulse-off phase.

The IED and ion angular distribution (IAD) are affected by the charge exchange collision within the plasma sheath. In the late 1990s, a mass-selected ion energy analyzer for RF electrodes was developed by Mizutani and Hayashi,\textsuperscript{125-127} and a bias frequency dependence of the actual IED was reported by Hikosaka et al.\textsuperscript{128} In 2000, Noda et al. reported on ion transport in HAR holes with a diameter of 200 nm in SiO\textsubscript{2} etching with a fluorocarbon plasma.\textsuperscript{129} The results showed that the ion flux and the ion energy were depressed at the bottom surface and strongly depended on the hole aspect ratio.\textsuperscript{129}

Recently, a low-frequency bias power, especially 400 kHz, has been used with HARC etchers. Typical process plasmas have an electron density on the order of 10\textsuperscript{11} cm\textsuperscript{-3}.\textsuperscript{130} By applying bias power to develop high-voltage sheaths, sheath widths increase to a range on the order of millimeters.\textsuperscript{130} As a result, the transit time of accelerating ion in the sheath is close to the half-period for a frequency of 400 kHz. Namely, since the ion travels with this half-period, the peak ion energy is maximized. When the same RF power is applied, the lower bias frequency of 400 kHz produces higher ion energies. Thus, a bias frequency of 400 kHz instead of the usual 2 MHz was implemented when applying the bias power. For low bias frequencies, ions transiting through the sheath in a small fraction of an RF period arrive with energy dependent on the phase and voltage of the bias when the ion enters the sheath.\textsuperscript{131} For high bias frequencies, ions transiting through the sheath over many RF periods arrive at the substrate with a narrower distribution centered about the average sheath potential.\textsuperscript{131} This motivated the use of multiple bias frequencies to customize the IEDs using a chirp, whose frequency was varied over time with an arbitrary waveform.\textsuperscript{131} Lamham and Kushner reported the computational results for biasing substrates using chirped frequencies in high-density, electronegative ICPs.\textsuperscript{131} The IED depended on the frequency range and chirp duration owing to transient shifts in the self-generated dc bias.\textsuperscript{131}

In addition, ions are transported from the plasma through the sheath to the surface. The IAD is actually spread over 5–10\textdegree around the essentially normal incidence angle. To investigate obliquely incident ions, Ui et al. reported that a bimodal IAD that consisted of ions obliquely incident mainly from two directions was controlled by modulation of RF power applied to four cathodes grouped with a phase shift. A nearly uniform IAD was obtained across the wafer.\textsuperscript{132} The anisotropic ions scatter at the mask edges and form specular reflections at the sidewalls of the mask and inside HAR features. However, no actual divergence of the IAD was measured.

To allow transport into deeper holes, advances in equipment have traditionally focused on increasing flow conductance and bias power. To achieve this, a new technology or approach is required for further advances.

This HAR etching trend is no longer viable as ARDE poses some critical limitations that cannot be overcome with the same approach. The transport of radicals and ions is the key issue in HAR etching. To enhance flux to the etching front, flow conductance was increased to maintain the purity of the bulk plasma with respect to the etching by-products. This increases the diffusion of new reactants into holes and removes by-products faster. Another way to enhance flux is to increase the temperature.

3.4 Cyclic and cryogenic processes for HAR patterning

In 1986, time-modulated (cyclic) etching was reported by Tsujimoto and Tachi.\textsuperscript{133,134} In 1988, cryogenic etching conducted at temperatures below 0 °C was proposed by Tsujimoto et al.\textsuperscript{135} To reduce the sidewall etching by both spontaneous radical reactions and ion-assisted reactions, a cyclic and cryogenic process was developed. Gotoh et al. reported that the estimation of IADs by fitting a cross-sectional profile to the simulated profile as parameterized by IADs versus the normalized etched depth.\textsuperscript{136} Under conditions of low pressure and high dissociations of both reaction byproducts and feedstock gases with a long residence time and high-density plasma, the mechanism of redeposition becomes complex. Tsujimoto et al. reported a high-flow etching apparatus that was equipped with a high-speed pumping system.\textsuperscript{134} They concluded that a cyclic and cryogenic process (with low pressure, high gas flow, and low temperature) was necessary, especially for n-type doped silicon anisotropic etching, due to thin sidewall protection.\textsuperscript{134} Kamto et al. reported the formation of through-silicon vias by ICP etching using SF\textsubscript{6}/O\textsubscript{2} gas chemistry at cryogenic temperatures.\textsuperscript{137} Control of the taper shape is critical in the optimization of the parameters for the gas flow rate, chamber pressure, ICP power, and substrate temperature.\textsuperscript{137} Therefore, chemical reactions with regard to both sidewall protection and ion-assisted reactions were indispensable for the realization of HAR features with high-volume manufacturability.

Recently, Parasuraman et al. reported successful patterning with an aspect ratio of 160 using a cyclic and cryogenic etching process for micro-electro-mechanical systems (MEMS) fabrication (Fig. 6).\textsuperscript{138} Conventionally, a fluorocarbon layer is used as a passivation layer to etch silicon
oxide selectively. Accordingly, to etch through this passivation layer, the ion energy must be high enough to penetrate the layer. Giving ions directionality by supplying RF power allows us to etch anisotropically. However, when the thermal energy of the substrate is very low (lower than the reaction threshold energy), etching does not happen spontaneously even without a passivation layer. In this situation, a very low ion energy is required to etch anisotropically. Based on this concept, etching at cryogenic temperatures was introduced to MEMS fabrication.

In addition to the cyclic process, a decrease in the treatment temperature enabled successful patterning etching at −110 °C with an aspect ratio of 125 as reported by Parasuraman et al. Unfortunately these technologies are far from ready for mass production in the semiconductor industry, even though they have already been introduced for MEMS devices. (Fig. 6) Solid lines represent the trend when only the width is reduced while the depth is maintained.

As shown in Fig. 6, Parasuraman et al. obtained empirical ARDE data that followed a semi-logarithmic law when a narrow feature width has a shallower depth. By extrapolating the trend of vanishing width, the ultimate aspect ratio for specific etching techniques were evaluated and the predictive ratio would maximize at around 3,000 because of processes free from aspect ratio dependence. Since a minimum thickness for the passivation layer is not included in the vanishing width, the experimental results prone to less than the predictive ratio.

### 3.5 Wet processes for HAR patterning

The wet process of metal-assisted catalyzed etching (MaCE) successfully fabricated HAR features in Si with an aspect ratio greater than 100. This process involves depositing a metal catalyst onto a Si substrate and immersing it in a bath containing hydrofluoric acid and an oxidizer such as hydrogen peroxide. The Si is only etched where the catalyzed metal contacts the surface by local oxidation of Si and etching of the oxide by etchants. The directional etching of Si is obtained by oxidation of the Si that is underneath the gold to an ionic form and the reduction of oxidants (H₂O₂), catalyzed by the gold, which produces holes (h⁺) in the Si. The product (H₂SiF₆) is soluble in HF. This process results in the removal of Si without any consumption of Au. The overall reaction for MaCE of Si with H₂O₂ and HF in a water solution catalyzed by Au can be written as

\[
\text{Si} + \text{H}_2\text{O}_2 + 6\text{HF} \rightarrow \text{Si} + \text{H}_2\text{O} + 6\text{HF} \rightarrow \text{Si} + \text{H}_2 + \text{H}_2\text{O} + 6\text{HF} + h^+ + 2e^- \rightarrow \text{SiF}_6^{2-} + 3\text{H}_2 \uparrow
\]

where H₂O₂ + 2H⁺ → 2H₂O + 2h⁺ is the reaction at the cathode (reduction) side and Si + 6HF + 2h⁺ → H₂SiF₆ + H₂ + h⁺ is the reaction at the anode (oxidation) side (Fig. 7). This implies that the etching direction is affected by the electron–hole concentration gradient in the charge carrier diffusion direction. Electric bias attenuates the etching in MaCE. As noted, the electrochemical etching reactions of Si are well known and the phenomenon was originally discovered in porous silicon formation by Uhlir in the 1950s. Si nanowire (NW) arrays were also fabricated by MaCE. Features in GaAs were successfully fabricated using potassium permanganate and sulfuric acid etchant solutions. For an InP substrate, hydrogen peroxide and sulfuric acid are used.

![Fig. 7.](https://example.com/fig7.png) (Color online) Simplified mechanism for metal-assisted chemical etching (MaCE) of silicon in HF and H₂O₂ mixed solution.

In 2005, Tsujino and Matsumura applied MaCE to form a cylindrical hole in Si with a diameter of 50 nm and a depth of more than 40 µm. They deposited Ag nanoparticles with diameters around 50 nm on a single-crystalline Si wafer and then immersed the wafer in a HF/H₂O₂ solution for 30 min. Tiberio et al. reported the fabrication of HAR structures (50 : 1; 25 nm width and 1.26 µm height) with vertically aligned gratings by MaCE with control of the electron–hole concentration. Very recently, Li et al. demonstrated HAR features (160 : 1; 130 nm width and 21 µm height) fabricated on Si by MaCE in a HF/H₂O₂ etchant solution using Ti/Au nanostructures as a catalyst. Furthermore, Li et al. reported the fabrication of HAR features (500 : 1; 16 nm width and 8 µm height) by a combination of MaCE and ALD processes. Pt catalyst patterns with a width of 16 nm were fabricated by Pt ALD on the sidewalls of microscale Si trenches. To realize MaCE in high-volume manufacturing, the patterning with metal catalysts, the removal of the catalysts, and the removal of contaminants must at least be considered. As a reference, MaCE may inspire the utilization of etching with controlled vertical directionality.

In this section, we reviewed processes for fabricating HAR features and controlling their CD uniformity. While the ALEt process has already been introduced, it still suffers from the aforementioned disadvantages. It is evident that conventional HARC technology will not be sufficient in the near future. A new approach will soon be needed to etch holes with aspect ratios greater than 100.

### 4. Boosting technologies for CD control at bottom of HAR features

#### 4.1 Self-limiting reactions to improve CD uniformity

Scaling from two to three dimensions and CD uniformity might be achieved by incorporating self-limiting reactions into fabrication processes. This emerging technology is required for the formation of contact holes with aspect ratios greater than 100, which remains harder than ever to achieve. It is evident that new approaches are required to continue the rate of advancement of previous years.

ALEt is a very challenging process, which offers improvements in selectivity and etching profiles but suffers from a
loss of throughput. Since the idea was suggested, many researchers have tried to improve both the throughput and etching profiles simultaneously. In recent years, the ALD process has developed dramatically but not to the same degree as ALD which was successfully introduced into production on the deposition side.

4.2 ALD of SiO₂, Si₃N₄, and SiC

The current state-of-the-art precursors, plasmas, and process conditions required to deposit conformal silicon dielectrics by plasma ALD are discussed here. Theoretical and experimental data will explain the observed reaction characteristics for the plasma ALD of SiO₂ and Si₃N₄, and the lack of success (so far) for SiC. Although plasma enables low-temperature deposition, it poses challenges in achieving isotropic film properties over the complex topography of today’s semiconductor devices.

For ALD of high-κ films, Hausmann et al. reported the ALD process of nanolaminates of Al₂O₃ and SiO₂ using trimethylaluminum (TMA), Me₃Al, and tris(tert-butoxy)isilanol [(Bu’O)₃SiOH]. They also reported smooth and conformal films of hafnium and zirconium oxides using six precursors of alkylamide — tetrakis(dimethylamido)zirconium [Zr(NMe₂)₄], tetrakis(ethylmethylamido)zirconium [Zr(NMeEt)₄], tetrakis(diethylamido)zirconium [Zr(NEt₂)₄], tetrakis(dimethylamido)hafnium [Hf(NMe₂)₄], tetrakis(ethylmethylamido)hafnium [Hf(NMeEt)₄], and tetrakis(diethylamido)hafnium [Hf(NEt₂)₄] — and water for oxidation. They pointed out that the alkylamide metal precursors were able to deposit smooth and conformal metal oxide films at temperatures as low as 50 °C. In contrast, diketonate and alkoxide precursors require a relatively high deposition temperature for precursor decomposition; however, the films were roughened by crystallization. Chloride and iodide precursors can be decomposed at temperatures as low as 180 °C; however, halogen impurities are incorporated into the films, and this cannot be prevented at low temperatures. Etching of the growing surface by the precursors themselves causes additional surface roughening and nonconformity.

As the dimensions of modern semiconductor devices continue to shrink below the current 14-nm technology node, novel processes for the deposition of highly conformal, low-temperature, and silicon-based dielectrics will be needed for applications that include sidewall spacers, barriers, and pattern layers. ALD is an ideal method for achieving high conformality and has been used in high volume manufacturing to deposit high-κ dielectric materials (e.g., HIO₂ and ZrO₂) for several technology generations.

Plasma-assisted ALD is the best-known method to meet low temperature (<500 °C) requirements, and it is currently used for depositing conformal silicon dielectrics such as SiO₂ and Si₃N₄. Typical dielectric ALD processing involves two discrete precursor steps: the first for the metal/silicon atoms and the second for the co-reactant (C, N, or O). These sequences are examined here in the context of the reaction mechanisms occurring during each step (Fig. 8). Silicon precursor selection is a daunting task, as nearly 10,000 silicon compounds are commercially available. These can be broadly grouped into six categories: amides, halides, alkoxides, hydrides, and alkyls (Table II). Predictions have been made that each ligand on a surface becomes replaced by a new ligand or nondirectional in nature, which means that they can collide with other ligands without changing the overall structure of the molecule.

When applying plasmas for silicon dielectric ALD, the other half-reaction in each ALD processing step typically involves oxidation, nitridation, or carbonization to form silicon oxide, silicon nitride, or silicon carbide, respectively. To simultaneously achieve good film properties at a low temperature, plasmas are required for the co-reactant step. Ovanesyan and co-workers reported a SiNₓ process employing a Si₂Cl₆ and NH₃ plasma at 400 °C as well as SiCNₓ processes employing a Si₂Cl₆ and CH₃NH₂ plasma at 400 °C and a SiCl₂(CH₃)₂ and NH₃ plasma at 350 °C. Faraz et al. reported a SiNₓ process employing a mono-aminosilane precursor, bis(tert-butylamino)silane (DBSAS) [SiH₂(N-sBu)₃], and N₂ plasma. Knoops et al. reported a SiNₓ process employing a bis(tert-butylamino)-silane (BTBAS) [SiH₂(NtBu)₃] precursor. Surface reactions during the ALD cycle were observed in situ by attenuated total reflection Fourier transform infrared spectroscopy (ATR-FTIR) with a ZnSe internal reflection prism. This technique enabled the in situ characterization of noninvasively probe-hindered intermediate species formed during the reactions. Surface –NH₂ and –NH species can be detected during Si–N–Si bond formation.

These precursors and N₂ plasma were used to apply deposits on HAR features, and the resulting growth per cycle (GPC) at the bottom was lower than that at the top. This implied that charge-free N radical species are isotropic or nondirectional in nature, which means that they can collide...
and recombine at the vertical surfaces of the trench nanostructures before they reach the bottom of the trench.\textsuperscript{156} Challenges arise in achieving high conformality at the HAR bottom with different chemical sources and plasma processes. Despite isotropic deposition on both sidewalls and horizontal surfaces, the anisotropic nature of ion bombardment preferentially improves film properties.\textsuperscript{156} Thus, a trade-off relationship between high conformality and high quality is observed for film deposition on 3D topographies.\textsuperscript{156} In addition, the surface roughness increases as the number of processing cycles increases. The rounded corners of deposited film and small amounts of impurities in the film composed of elements of the processing gases were studied experimentally and computationally in ALD deposition of SiN film using hexachlorodisilane and hydrazine.\textsuperscript{173} They speculated that plasma could generate undercoordinated surface sites such as nitrogen dangling bonds, and subsequent precursor adsorption occurred on uniform surface covered with both –Si–H and silyl (–SiH\textsubscript{3}) groups. The resultant –Si–H surface may rapidly react, as compared with the –SiH\textsubscript{3} site. This rate-difference causes to remain unreacted bonds and to give rise non-conformality. In addition, during plasma step, the fragmented ligand species could be redeposited as impurities on the growing film surface.\textsuperscript{157} To solve the problems of conformity and quality, stringent control of ion and radical transport inside HAR features by controlling the plasma sheet is challenging.

### 4.3 Self-limiting surface reactions for atomic-level control in Al\textsubscript{2}O\textsubscript{3} ALD

The factors that lead to self-limiting surface reactions are outlined here. The defining characteristic of ALD is reviewed using experimental data from the literature. Ligands can persist at low temperatures, as explained via computed activation energies that show the so-called cooperative effect.

A reaction mechanism for oxygen plasma was also computed. One of the attractive features of ALD is that it works at much lower temperatures (e.g., 150–300 °C) than comparable chemical vapor deposition processes, and this enables deposition onto thermally sensitive substrates. In some cases, ALD can be achieved at even lower temperatures, but such temperatures often require plasma, as the thermal process may have a too low growth rate to be useful. The prototypical ALD process of TMA and water is one such example. In this process, the amount of alumina deposited in a ALD cycle drops steadily as the growth temperature is reduced below 200 °C, whereas the growth rate is maintained even down to room temperature if TMA+O\textsubscript{2} plasma is used. It has long been supposed that the reason for this lies in thermal activation being required for some reaction step within the overall ALD process. Recent experiments and calculations have revealed which reaction step is involved, and this has modified our view of ALD surface chemistry. Experimental results from the Eindhoven University of Technology detected persistent methyl groups at the end of the H\textsubscript{2}O pulse in low-temperature alumina ALD,\textsuperscript{174} and Stanford University obtained analogous results for zinc and tin oxide ALD.\textsuperscript{175}

Now, we present an overview of computed reaction mechanisms for thermal ALD. Specific surface reactions were previously computed to show coverage-dependent kinetics for alumina and hafnia ALD and to identify the types of species that can consequently be expected to persist at the surface.\textsuperscript{176} Computed activation energies at metal oxide surfaces in different local environments show the central role played by the coordination number. Reaction from the anchor OH group to threefold coordinated Al(CH\textsubscript{3})\textsubscript{3} was in computation estimated an activation energy of 0.28 eV. Reaction from the other neighboring O to one- or two-coordinated Al(CH\textsubscript{3})\textsubscript{3} was higher activation energy of 0.74 eV, where the Al–C bond breaks by the proton transfer to the methyl group.\textsuperscript{176} By computing complex surface models, the co-adsorption of precursors leads to a lower activation energy than previously reported. In some cases, these are the first reaction pathways computed to be viable at process temperatures. Undercoordinated oxygen is reactive towards the adsorption of Lewis acidic metal precursors, such as Al(CH\textsubscript{3})\textsubscript{3} and Hf(N(CH\textsubscript{3})\textsubscript{2})\textsubscript{4}, which in turn through the cooperative effect make hydroxyl groups more Brunsted acidic. A complementary set of reactions is observed when a ligand-covered surface is exposed to H\textsubscript{2}O, with co-adsorption activating both H\textsubscript{2}O towards dissociation and ligands towards protonation and desorption. These findings show how saturation of the coordination number can explain the self-limiting nature of ALD, from which the unique advantages of this technique are derived. These effects may be even more acute in the case of the thermal ALD of nitrides with the less reactive co-reagent NH\textsubscript{3}.

Comparison between the oxygen plasma and thermal H\textsubscript{2}O mechanisms shows a different mechanism for TMA+O\textsubscript{2}–plasma from that predicted by density functional theory. A range of gas-phase oxidation by-products are predicted by the calculations, with CH\textsubscript{2}O predominating. The prediction confirms that each ligand on the surface becomes replaced through oxidation by a hydroxyl group, independent of the by-product.\textsuperscript{177} The factors leading to the densification of newly deposited Al atoms into bulklike coordinated structures are discussed, along with the consequences for the morphology of the film. Although the oxidation reactions of the plasma are initially spontaneous, they rapidly become self-limiting because of the limited reducing ability of the methyl-covered surface during ALD.

### 4.4 State-of-the-art of ALEt

Owing to the advantage of self-limiting surface reactions to achieve etching uniformity, ALEt has received much attention because it provides atomic-level control and material selectivity. The ALEt technologies are classified into anisotropic and isotropic etching (Fig. 9). The former classification includes digital etching, which is conducted in separate steps for etchant adsorption and removal of the generated volatile products after chemical reactions.\textsuperscript{178,179}
The anisotropic nature is obtained by vertical ion incidence to enhanced ion-assisted chemical reactions. Thus, this includes cyclical processes, which involve sidewall protection and ion-assisted reactions. Recently, the IED has been reconsidered for the optimization of ion-assisted reactions. A very narrow IED is provided by a beam etching system, in which ions are extracted by grid-type single or multiple electrodes. The use of RF bias control for ion-energy control gives a relatively broad IED or multiple peaks in the IED. Penetration depths of ions into the bombarded materials are determined statistically for their incidence energy; thus, a narrow IED band is desirable for controlling reactions at the atomic level. Comprehensive reviews are found elsewhere.\(^{180-183}\) Tabata et al. reported that a quasi-ALEt technique that controls radical flux and ion flux independently could resolve etching interruption (etch-stop) issues due to the reduction of excess deposition on small area in contact-hole of SiO\(_2\), when a TiN hard mask was conventionally used as occurred etch stop.\(^{184}\) Tsutsumi et al. reported the ALEt of SiO\(_2\) films with cyclic fluorocarbon deposition and O\(_2\) plasma removal.\(^{185}\) Ishii et al. reported a similar ALEt with a cyclic process of CHF-gas chemistry and Ar irradiation that improved material selectivity of SiO\(_2\) over Si.\(^{186}\) In addition, the directional ALEt of GaN and AlGaN using cyclic Cl\(_2\) plasma chemisorption and Ar ion removal was reported under conditions of low ion energy ranging from 50 to 100 eV within the self-limiting regime, as reported by Ohba et al.\(^{187}\)

In contrast, isotropic etching includes a sequential process for ammonium-salt-based modified layer formation and thermal evaporation of the salt, as originally reported for SiO\(_2\) removal by Nishino et al.\(^{188}\) The salt is essentially formed by a self-limiting reaction and remains on the reacted surface. Then, it is desorbed by thermal heating without leaving any residue on the surface. The isotropic and self-limiting properties are useful for conformality control in HAR features. Recently, Shinoda et al. reported the thermal cyclic etching of SiN, TiN, and W films developed on the basis of salt formation and thermal evaporation.\(^{189-192}\) Kofuji et al. reported that a lateral sidewall of W was etched uniformly using NF\(_3\) plasma chemistry.\(^{193}\) In other words, the adsorption of reactants represents the processes of physisorption and chemisorption on the surface, and then complex salts form to bind the reactant as ligands with atoms that exist on the surface. If the complex formation is self-limiting and material-selective, and the product is volatile, then isotropic ALEt is enabled. A very important merit of plasma processing is that it enables the deposition and etching of materials with high quality at low temperatures through the use of energetic species present within the discharge, especially from an industrial perspective.\(^{194}\) Plasma-enhanced ALEt at low temperatures has enabled self-aligned patterning, widely considered a breakthrough technology for improving CD uniformity and the downsizing of devices. Honda et al. extensively addressed self-limiting processes, such as ALEt, which adopted the advantage of the independence of the pattern density and the location on the wafer.\(^{195}\) Namely, the atomic layer processes ALD and ALEt are free from aspect-ratio dependence due to their self-limiting nature. This property is indispensable for high-volume manufacturing.

5. Co-optimization among design, system, process, and technology for manufacturability

5.1 Basics of logic circuit design for fin-FETs

Currently, ultralarge scale integrated circuits (ULSIs) use a discrete CMOS transistor as a circuit element, which consumes less power than its bipolar counterparts.\(^{196}\) The CMOS logic involves combinational circuits, whose outputs depend only on the current inputs, and sequential circuits, whose outputs depend on both current inputs and previous input, with registers as memory.\(^{195}\) The physical implementation of the logic gates and the registers from transistors has a major impact on performance, power, and cost.\(^{195}\) Previously, Mead and Conway popularized scalable design rules based on a single parameter \(\lambda\) that characterizes the resolution of the process.\(^{195,196}\) Thereafter, a physical layout can be automatically generated that has a simple layout style based on a “line of diffusion” rule commonly used for standard cells in automated layout systems.\(^{195}\) Note that this style consists of four horizontal strips: metal ground at the bottom of the cell, strips for n-diffusion and p-diffusion, and a metal power supply at the top. The power and ground lines are often called supply rails.\(^{195}\) In contemporary standard cells, polysilicon is generally not used as a routing layer, so the cell must allow second-metal-layer to first-metal-layer (metal2-to-metal1) and first metal layer to polysilicon (metal1-to-poly) contacts for each gate. While this increases the size of the cell, it allows free access to all terminals in the metal routing (Mx) layers.\(^{195}\) In the standard cell design flow, the logic design is created by a synthesis tool at the register transfer language (RTL) level. The synthesis tool maps the algorithmic description of the design into a collection of standardized Boolean logic functions expressed as NAND and NOR gates, inverters, latches, and other elements. Then the physical layout is rendered by an electronic design automation (EDA) tool and a standard cell library to build the appropriate place and route (PnR) solution.\(^{197}\) The transistor structure moves from a planar to fin-FET structure with processes such as multiple patterning, self-aligned double patterning (SADP), and LELE and cuts. For this, the fin-FET layout design is also modified.

A major difference in the fin-FET design is the use of triple-gate (tri-gate) and double-gate structures.\(^{198}\) In a short-gate (SG) structure, the left and right sides are connected together, but in an independent-gate (IG) structure, the top part of the wrap-around gate structure is etched out.\(^{198}\) The fin-FET performance is affected by the cross-sectional area of the fin and the fin shape.\(^{198,199}\) A triangular fin can reduce leakage current.\(^{198,199}\) A rectangular fin has better short channel effects metrics, in particular, a sub-threshold slope, gate-induced-drain leakage (GIDL), and drain-induced barrier lowering (DIBL). These are caused by strengthening the electrostatic field without the coupling of field lines due to the small dimensions. Thus, increasing the fin heights directly improves performance, such as current drivability, and fin height has replaced width as the conventional design parameter.

Standard cells are defined by dimensions such as gate pitch (GP), contacted poly pitch (CPP), metal pitch (MP), cell height, and fin pitch (FP), as shown in Fig. 10.\(^{200}\) The MP describes the smallest width and space combination allowed.
The parasitic capacitances in fin-FETs have been analyzed in a number of studies.\textsuperscript{205} A multigate fin-FET structure requires significant attention to the parasitic capacitances, that is, fringe capacitances by fin sidewalls, associated with 3D fins, multigates, dummy gates, and trench contacts.\textsuperscript{205} The fringe capacitances are calculated by the addition of gate-to-contact, inner-gate-to-fin, and outer-gate-to-fin components.\textsuperscript{206} The parallel plate capacitance is $S/D$ diffusion along the gate sidewall and fin tops.\textsuperscript{206} The overlap capacitance originates from an overlap between the $S/D$ region and the gate.\textsuperscript{206}

The parasitic capacitance between the gate and contact is a very important contributor to the power-performance trade-off and is mitigated by very low $k$ materials or partial airgap spacer technology.\textsuperscript{207}

The contact resistivity component is also very important and is mitigated by significant increases in the source and drain surface doping and increasing the effective mass of the contact metal to make the contact quasi-ohmic and reduce interface carrier scattering.\textsuperscript{208}

In metal gate integration (RMG), the gate length requires very thin barrier metals with a thickness thinner than 2 or 3 nm and multiple threshold voltage ($V_t$) schemes that do not involve increases in barrier thickness.\textsuperscript{209}

Although patterning solutions exist, crossing the 40 nm CPP barrier is a daunting proposition from economic, yield, device performance and process control standpoints and quite possibly may not happen.

5.3 Cell-level: Circuit density scaling and DTCO
The greatest challenge in modern very large scale integration (VLSI) design is not in designing the individual transistors but rather in managing the system complexity.\textsuperscript{195} Digital VLSI design is often partitioned into five levels of abstraction: architecture design, microarchitecture design, logic design, circuit design, and physical design.\textsuperscript{195} To deal with these interdependences, microarchitecture, logic, circuit, and physical design must occur in parallel to some extent.\textsuperscript{195}

DFM is a methodology for ensuring that a product can be manufactured repeatedly, consistently, reliably, and cost-effectively by taking all the measures necessary, starting at the concept stage of a design and implementing these measures throughout the design, manufacturing, and assembly processes.\textsuperscript{210}

The bleak prospect that device CD downsizing will stop is alleviated by the fact that circuit area scaling is enabled by several constructs (scaling boosters) at the circuit level that allow the reduction of the standard cell size through a reduction of the number of metal tracks. The power performance trade-offs in the implementation of such constructs constitute, in essence, a reason for DTCO. Examples of scaling boosters include single-diffusion breaks, self-aligned gate contacts, fin-cutting sequences (metal cutting), super-vias (for dense Mx routing), self-aligned blocks, and fully aligned vias.

Diffusion breaks refer to the space separation between two active device regions, where two dummy gates separate the two active regions, and the diffusion area is the silicon (non-isolation) region.\textsuperscript{197} This provides a process window to tolerate CD variation in shallow trench isolation (STI) and gate-to-STI misplacement; however, the STI dimension needs to be very narrow so that the fin ends of adjoining...
fin-FETs can tuck under one narrow dummy gate, even under worst-case misalignment.197 Patterning and etching such a narrow STI are very difficult and the narrow STI CD can lead to high device leakage due to poor isolation.197

Self-aligned contact (SAC) of gates and the S/D is also challenging. In the older planar technology, the gate pitch was relaxed such that S/D contacts and gate contacts could easily be placed next to each other without causing any shorting risk.197 SAC mitigates the issue of S/D contact to gate shorts by fully encapsulated the gate metal with a dielectric spacer and gate cap.197

The “fin cut first” and “fin cut last” schemes differ in the sequencing of dummy fin removal.197 The “fin cut first” scheme defines the cut on the fin hard-mask level such that only one RIE step is needed to form the final fin shape. This sequence makes the isolation process simpler by requiring only a single oxide void-free “gap fill” and planarization.197 Issues such as the complexity of the cut etching, the pattern-dependence of dielectric deposition, and the difficulty of chemical–mechanical polishing (CMP) arise from the non-uniformity of the pattern density. The “fin cut last” scheme, as the name implies, removes the dummy fins after the final fin image is etched into the substrate. This means that the fin has to be cut after the isolation process, which requires an additional isolation process after the fin cut. Therefore, the fin cut last scheme requires a larger number of process steps than the “fin cut first” scheme. As the fin is cut across the full topography of the final fin structure in the “fin cut last” scheme, cut mask misalignment can lead to spikes in the residual dummy fins.197

Downsizing of ICs requires dimensional scaling from previous nodes for standard cell and custom logic requirements, and high-yield, low-cost integration schemes.211 The reduction of cell area can be achieved by fully self-aligned gate contact, which makes the gate contacts over the active region, instead of conventional placement of gate contact outside of the diffusion area.211 The self-aligning technology can be introduced at all levels, such as the front-end-of-line (FEOL), middle-of-line (MOL), and back-end-of-line (BEOL). Sherazi et al. reported boosting technology for a low-track-height standard cell design in the N7 node.211 Some of these constructs are needed to scale the number of metal tracks per standard cell down to five or even four, as shown in Fig. 11. More innovative schemes such as buried power rails can be powerful enablers of area scaling without decreasing the CPP.

### 5.4 STCO for merging cross-point devices for logic and non-volatile memory

The “one device fits all needs” era is coming to an end, and a clear contender for replacing today’s Si-based CMOS technology seems to elude us. In the longer term (10–15 years), we believe that the most promising progress in increasing functionality per unit cost will be through STCO for maximum system performance.

A recent example is Imec’s resistive random access memory (RRAM)-based synaptic processing unit212 optimized for machine learning applications. Through smart system-level partitioning, one can optimize system components with the most suitable technology. The Imec system was developed through low-cost 3D technologies and monolithic 3D integration. A variety of new material-based devices also have the potential to increase functionality at the system level by co-integration in the BEOL processing to fabricate interconnects with standard Si-based technology.

More disruptive technologies, such as the proposed nanofabrics, may allow the STCO-driven hybrid (functional) scaling trend to continue. The volume requirements in application domains will provide economies of scale that will justify this hybrid scaling approach.

The traditional standard-cell based design methodologies rely on a large library of standard cells to synthesize the logic of the design. As the cells are placed next to each other, a large number of layout patterns are created as each abutment of two cells has the potential to generate a new pattern neighborhood at the cell boundary. Although regular design methods are being adopted in the industry, the practices defined so far do not sufficiently restrict the number of layout patterns, especially at the edges of cells. Such macro-regularity for logic can only be enabled by carefully designing the edges of all cells. The regular design fabric and templates, which are introduced in the next subsection, specify the allowed layout constructs and layout neighborhoods using a completely prescriptive approach to IC layout design.213 It is not easy to enforce all the pattern constraints desired by the regular design fabric by providing a grid-based design infrastructure to designers.213 To ensure macro-regularity and enable efficient designs using regular design fabrics, we propose to add a level of abstraction between the standard cells and the regular design fabric, which we call logic templates.213

Also, any memory architecture emerging to displace the NAND and DRAM incumbents must have a similar feature density. The tremendous number of features on a wafer also makes it critical that the statistical process capability also scales with each node.

New memory technologies, such as 3D cross-point (3DXP), also require the application of pitch multiplication and HAR etching technology to enable the targeted bit densities and high manufacturing volumes. These new memory technologies, similarly to 3DXP, also have unique dry-etching process challenges.

### 5.5 New materials, novel devices, and novel device constructs

One key enabler of the scaled gate length is the NW-based
device architecture\textsuperscript{14} that allows optimal gate electrostatic control; a less electrostatically effective variant is the nanosheet (NS)-based device architecture, although the NS device architecture has the benefit of increased drive currents per unit area.\textsuperscript{215} In addition, downsizing the CPP beyond 40 nm would require either a transition to EUV-based SADP or self-aligned quadruple patterning (SAQP), an increase in the number of NWs or NSs in the device stack, and extreme parasitic load mitigation solutions,\textsuperscript{216} including the implementation of a full airgap spacer.

A steep on/off current transfer curve is required. High-mobility materials (Ge and III/V) have been hailed for years as candidates for replacing Si in scaled devices. However, advances in Si technology, the advent of NW and NS devices, and the many issues that plague high-mobility materials (leakage, defects, low temperature budgets, and cost) make their introduction less and less likely. A SiGe p-channel FET (pFET) or the introduction of stress in the transport channel is required. High-gain transfer characteristic is possible for fin-FET devices,\textsuperscript{217} however, their use is questionable in the NW/NS context.

Moving beyond traditional fin-FET devices, we see vertical FETs (VFETs—where the device channel is vertical) or complementary FETs (CFETs—where NW-based P and N devices are stacked on top of each other) as contenders to extend circuit area scaling (Fig. 12).\textsuperscript{218}

These devices are being built using mostly existing semiconductor technology (based on Si or SiGe). Although some innovations, such as nanotubes, are required, so they allow continued scaling through extreme gate shrinking. For example, VFET devices can be the most effective way of scaling static random access memory (SRAM), and III/V materials can provide the best figures of merit for the analog parts of a chip. By co-integrating these components in the most cost-effective fashion with standard Si-based logic, we can optimize the system functionality.

The 2D-material-based devices and spin-based devices also have the potential to increase functionality at the system level. Due to the employment of low-temperature fabrication, such 2D-material-based devices can be inserted in standard BEOL processing, and they offer the potential for stacking in a layered fashion.\textsuperscript{219} Another class of devices involving spin propagation is also a potential candidate for co-integration with standard logic.\textsuperscript{220}

6. Future challenges

We stress that atomic layer processes will continue to improve the CD uniformity at the bottom of HAR features. In 1992, Gottscho et al. discussed macroscopic CD uniformity. At that time, RIE lag resulted from microscopic transport phenomena within a single feature while microloading referred to a local dependence of the etching rate on the pattern density for identical features.\textsuperscript{89} The microloading is caused by the same mechanisms as those cause macroloading in one type of ARDE.\textsuperscript{221} The ARDE mechanisms were categorized into (1) Knudsen transport of neutrals, (2) ion shading, (3) neutral shadowing, (4) differential charging of an insulating microstructure, (5) field curvature near the conductive topography, (6) surface diffusion, (7) bulk diffusion, and (8) image force deflection. Fujiwara et al. pointed out that the etching rate was influenced by the transport of ions and neutrals inside HAR features.\textsuperscript{35} Knudsen transport provides a simple description of the neutral transport that occurs in HAR features; that is, the neutrals travel without colliding with reflections at the wall, giving them a cosine angular distribution. As a result, the neutral transport rate is obtained by molecular diffusion and the surface sticking coefficient (versely related to the scattering coefficient). Accordingly, neutrals with thermal motion having randomly directed incidence at the top entrance to HAR features have a net flux into the feature with thermal velocity given by

\[
\nu_{\text{in}} = \sqrt{\frac{8kT}{\pi m}},
\]

where \(k\) is Boltzmann’s constant, \(T\) is the temperature, and \(m\) is the atomic or molecular mass. Then, during the transport, the sticking of neutrals determines the neutral flux to the bottom of HAR features. Also, bulk or surface diffusion of the sticking neutrals may need to be taken into consideration. Briefly, the reduced sticking of neutrals at the sidewalls of HAR features results in a higher flux of neutrals that reach the bottom. The ARDE properties are crucial for understanding the effects of gas and surface temperatures. For spontaneous etching regimes, sidewall passivation is considered by exploring low-temperature processes. In contrast, neutral starvation may possibly be solved by high-temperature processes. Nevertheless, the reaction probability for the targeted etching reaction and the net transport of reactants reaching the bottom determine the etching rates and ARDE characteristics. However, we emphasize again that self-limiting processes take advantage of the independence of macroscopic and microscopic loading effects. Much work remains for developing revolutionary methods that overcome these problems.

Returning to the starting point, we reconsider the surface etching reactions occurring at an HAR hole bottom (Fig. 13). In accordance with RIE behavior, radicals and ions synergistically contribute to the formation of etching products. Ohiwa et al. argued mechanisms of etching rates for bottom on reflections at hole sidewall during etching with ion transport inside HAR\textsuperscript{106}. As discussed in Sect. 3.3, the IED and IAD with higher bias power is becoming a critical technology and a new boosting technology is necessary.

For understanding of the atomic-level processes, information about the details of these processes remains insufficient,
because dynamical changes in etching state under determined by a stochastic process that a future state depended on the present state, i.e., non Markovian-process. To control the etching process as a system, final state should be causal deterministic from initial states. Namely, of great interest is the nonequilibrium dynamics of the etching processes. Similarly consider the dynamics as a system, a response time with or without a relaxation determines characteristic frequency. Note that self-limiting processes become free from this frequency and robust in space and time. In a essence meaning, the atomic-level and self-limiting reactions constitute on taking advantages for improving any uniformity, namely, achieving a temporal- and spatial-variation-free (adaptive) control. A remained main problem is how to know a present state of the etching system. Together with actual observations of behaviors of the system variables at real time, a mathematical description of the dynamics may solve strictly or probabilistic causal states that represents, for example, etching rates, etched profiles, CD uniformities, and surface sticking coefficient.

Currently, the IEDs and IADs are obtained by a consensus of the importance of CD uniformity at a HAR trench or hole bottom. The Kushier group reported computational results for IEDs and IADs in plasma etchers operated at multiple frequencies. However, charging effects at the HAR feature bottom are very complicated. Charge build-up inside insulating HAR features may distort the ion trajectories, causing ions to strike the sidewalls. This is essentially caused by electron shading of the HAR features because the electron flux is isotropic. To solve this problem, the anisotropic flux of electrons or negatively charged ions should be considered. Furthermore, when charges between electrons and ions are balanced at the bottom, an electropositive potential develops there. Economou and coworkers reported electrostatic lenses that were fabricated by constructing a surface metal plate and substrate and these lenses were achieved by electrostatic focusing of ions inside HAR feature sandwiched two electrostatic electrodes, realizing the patterning of nano-features, they called as nanopantography. Their collimating method with a monoenergetic ion-beam has attracted great attention also for the purpose of IED and IAD control. The resultant ion flux needs to increase at the wafer-biasing voltages. Further study of the control of CD uniformity at the HAR feature bottoms is still required with an emphasis on charging, absorbates, and redeposition of by-products.

We presented some state-of-the-art examples to demonstrate that progress can be made by coordination among theory, computation, and experiments (Fig. 14). To do this, fundamental data are indeed necessary. For ion-induced reactions on an etched surface, the etching yield per incident ion is basic information that we need to have. In recent research, Karahashi et al. reported the experimental etching yield of amorphous carbon by fluorocarbon ion irradiation. Kuboi et al. demonstrated a coordinated surface reaction model based on reaction probabilities for ion-induced reactions. Recent advances in computational power are becoming useful for performing virtual experiments. Briggs et al. reported that photoionization of molecular nitrogen and methane in elementary electron-molecule collisions was studied theoretically by the molecular R-matrix method and confirmed experimentally with good agreement. Hayashi et al. investigated the electronic properties of HBr, Cl2, and O2 molecules using computational chemistry and reported that electron-attracted negative ions of these compounds tend to dissociate into atoms and negative ions, except O2 (H+Br−, Cl+Cl−, and O2−), and ionized positive ions form molecular ions in the threshold...
reaction-chamber.239 phase reaction-chemistry and plasma physics inside the and experimental approaches. The levels comprise of gas-pro
searching of a new recipe and applications of the optimiza-
on plasma-processed surface and etching yield of target (Fig. 14). An etching recipe that is conducting in the chamber conditions to
elucidate. Arti
mechanisms of atomic layer reactions are challenging to
exploit an interdisciplinary environment involving physics experiments. To realize this approach, we need to practically
con
bottom of features. We also outlined the major challenges for
macroscopic CD uniformity, and (3) CD uniformity at the
HAR etching technologies with respect to (1) CD scaling, (2)
We have discussed recent advances in fabrication technolo-
original content from this work may be used under the terms of the
Creative Commons Attribution 4.0 licence.
Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
