PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Addressing Future Analytical Requirements of Electronic Materials

Badih El-Kareh

Phil. Trans. R. Soc. Lond. A 1996 **354**, 2597-2617
doi: 10.1098/rsta.1996.0118

Addressing future analytical requirements of electronic materials

By Badih El-Kareh

Semiconductor Research and Development Center, IBM Corporation, Hopewell Junction, New York 12533, USA

The drive toward faster, denser, cheaper, lower power consuming and more reliable semiconductor products is predicted to continue relentlessly toward the end of the next decade, with a new generation introduced every three years. This is achieved by reducing horizontal and vertical device geometries, and by introducing more sophisticated device structures. While these developments move at a fast pace, industries focus on rapid yield learning to reduce development costs. The key to fast process control and defect reduction is the implementation of efficient analytical and simulation tools. After discussing trends in memory and logic, this paper focuses on key elements of CMOS structures and on characterization needs for successful development of future technologies.

1. Introduction

The development cycle of a semiconductor product begins with the choice of a set of manufacturing tools that will be available during the production stage. Process and device parameters are then specified, consistent with the tool capabilities, to optimize circuit performance and density, yield, cost and reliability. Initially, engineers rely heavily on a sophisticated array of process, device, and circuit simulation tools to evaluate technology options. Past generations have shown, however, that process modelling alone cannot accurately predict vertical and horizontal geometries that result in the desired device and circuit parameters. For some new structures a satisfactory model may not even be available. While taking advantage of process simulation to narrow down technology options and to reduce the number of costly experiments, engineers rely heavily on analytical tools, processing knowledge and experience and self-consistency checks to arrive at the desired profiles (El-Kareh 1994). Therefore, considerable process characterization is required during the development stage to 'centre' the technology, identify yield detractors and to calibrate process models. Even as the process matures, analytical techniques remain important to tighten tolerances, improve circuit performance, reduce die size and investigate process simplifications to reduce cost. This cycle repeats every generation when the minimum printable feature size is reduced and new technologies are introduced to improve density and performance.

One must be very careful when making forecasts and defining fundamental device limits. For example, in 1979 the minimum 'practical' MOSFET channel length and gate oxide thickness were predicted not to exceed, respectively, 0.5 μm and 9.4 nm at 1.5 V because of field limitations (Masuda *et al.* 1979). Today, products at 0.20 μm

Phil. Trans. R. Soc. Lond. A (1996) **354**, 2597–2617
Printed in Great Britain

Table 1. *Suggested technology roadmap*

(Adapted from the 1994 National Technology Roadmap for Semiconducters (Semiconductor Industry Association).)

year	1995	1998	2001	2004	2007	2010
min. feature (μm)	0.35	0.25	0.18	0.13	0.10	0.07
power supply voltage (V)	3.3/2.5	2.5/2.5	1.8/1.8	1.5/0.9	1.2/0.9	0.9/0.9
max. wafer diameter (mm)	200	200	300	300	400	400
isolation	LOCOS/STI	LOCOS/STI	STI/SOI	STI/SOI	STI/SOI	STI/SOI
S/D extension depth (nm)	70–100	50–120	30–80	20–60	15–45	10–30
gate oxide equivalent (nm)	10	6	4.5	4	< 4	< 4
max. fault density (m^{-2})	240	160	140	120	100	25
max. contact aspect ratio	4.5:1	5.5:1	6.3:1	7.5:1	9.0:1	10.5:1
heavy metals ($\times 10^{10}$ cm^{-2})	5.0	2.5	1.0	0.5	0.25	< 0.25
DRAM masking steps	20	22	22	24	24	26
DRAM cell size (μm^2)	1.50	0.60	0.24	0.10	0.04	0.015
DRAM bits/die	64 M	256 M	1 G	4 G	16 G	64 G
DRAM die size (mm^2)	190	280	420	640	960	1400
max. μP transistors cm^{-2}	4 M	7 M	13 M	25 M	50 M	90 M
max. μP size (mm^2)	250	300	360	430	520	620
max. ASIC die size (mm^2)	450	660	750	900	1100	1400
max. # wiring levels	4–5	5	5–6	6	6–7	7–8
max. H. perf. power (W)	80	100	120	140	160	180
max. battery power (W)	2.5	2.5	3.0	3.5	4.0	4.5

channel length and 6 nm oxide thickness that operate at a nominal power supply voltage of 2.5 V are entering the final stage of development. Also, it was predicted in 1979 that photolithography will 'never break the 1 μm barrier'. Today, deep UV, phase-shift masks and off-axis illumination have extended photolithography capabilities to subquarter micron features. There is no reason to believe that trends seen in the past years will not continue toward the next decade. The minimum printable feature size is predicted to continue to decrease by a factor of approximately 0.7 every three years, resulting in a corresponding increase in memory density by a factor of four and in logic circuit density by a factor of two to three (Masuda *et al.* 1979; Hu 1994; Singer 1992). These predictions are summarized in table 1 and figures 1 and 2.

In each generation attention focuses on lithography tools and their ability to reliably reproduce features at deep submicron dimensions. Shrinking horizontal geometries, however, is only part of the challenge when introducing a new generation. To optimize performance, yield and reliability, the power supply voltage and gate oxide thickness must be reduced, the vertical and horizontal impurity profiles must be readjusted and, for some applications, a new isolation scheme and interconnect technology must be introduced. The reduction in horizontal and vertical dimensions brings with it an increased sensitivity to process deviations that poses more stringent

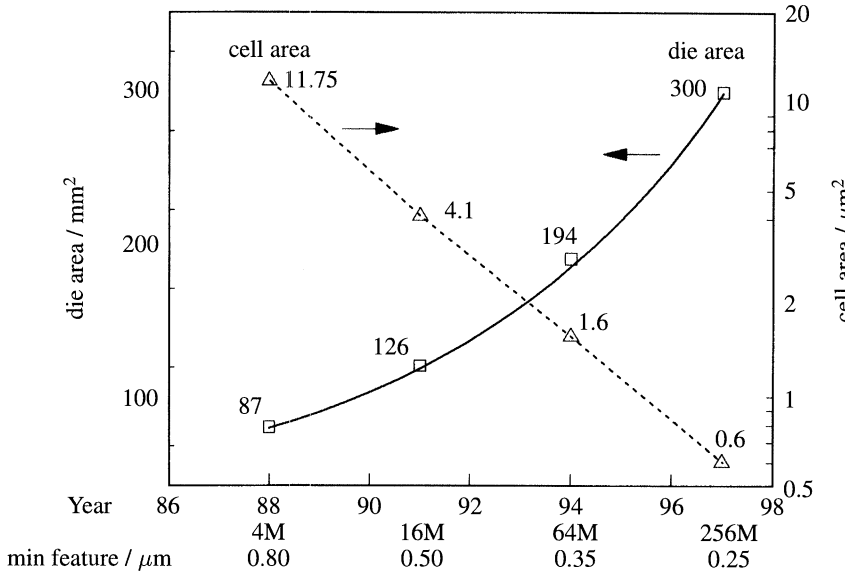

Figure 1. DRAM trends.

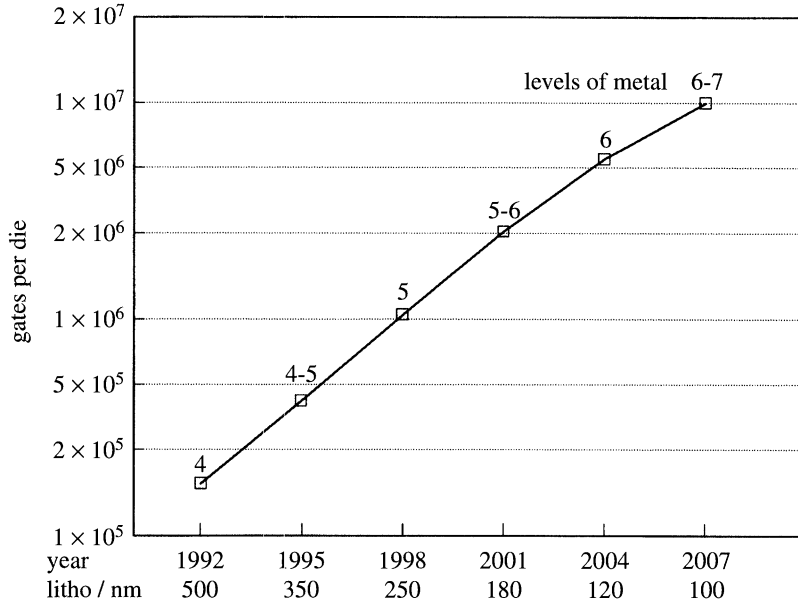

Figure 2. Progression of logic technology and density.

demands on analytical tools with respect to sensitivity, accuracy, spatial resolution and repeatability. Also, for a product to be cost competitive, the period from research and development to full production must be reduced every generation (figure 3). This requires shorter turnaround times for test and characterization, real-time feedback and *in situ* rather than offline process monitoring.

Complementary metal oxide semiconductor (CMOS) is the leading technology for memory, logic and analogue designs. It has captured over three-quarters of the world-wide semiconductor market because of its low power density, good immunity to noise

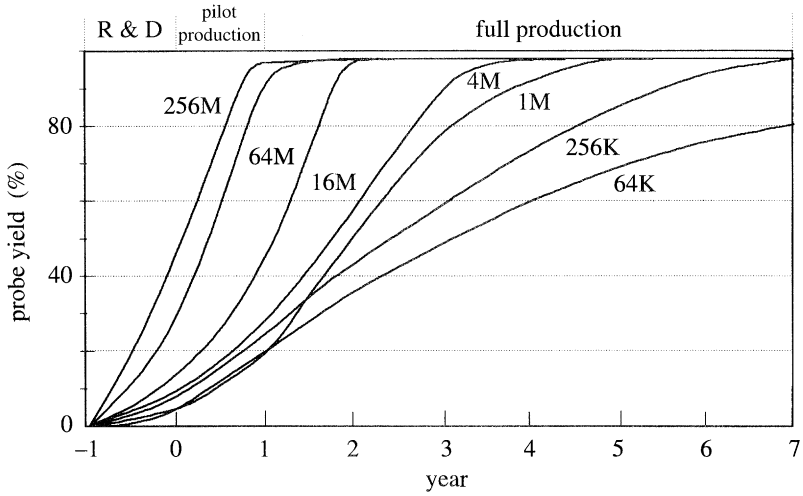

Figure 3. Industry average probe yield versus time (DRAMs, measured after repairs). Source: VLSI Research Inc. (see Hutcheson 1995).

and single-event upsets, easy scaling and compatibility with integrated systems on a chip. This paper will therefore focus on key CMOS technology and device features. The analysis is, however, also applicable to other technologies and to materials, such as bipolars, BiCMOS and compound semiconductors. Important MOSFET characteristics and their relations to process parameters are first discussed. This is followed by a discussion of CMOS technology features with a focus on memory and logic. Analytical requirements for successful development of future technologies are then addressed.

2. MOSFET parameters

Figure 4 shows a top view and cross sections of an n-channel and a p-channel MOSFET. While the basic principles of MOSFET operation do not change, the complexity of the structure increases with each generation, as can be seen from table 2 and figures 5–8.

(*a*) *Source and drain*

The most important source/drain properties are the lateral extent, 2D–3D field, series and contact resistances, junction and overlap capacitances, junction leakage and breakdown voltage. These properties are tightly related to the lateral and vertical junction profiles.

(*b*) *Effective channel dimensions*

The region under the thin gate dielectric is called the channel. There is typically a difference between the designed channel length and width, L and W, and the electrically effective dimensions, L_{eff} and W_{eff}. These differences are caused by lithography and etch 'biases', and by the lateral scattering/diffusion of dopants. Variations in L_{eff} and W_{eff} can be attributed to process tolerances. The interpretation of L_{eff} becomes, however, more ambiguous as the channel length is reduced, requiring accurate profile measurements for its definition.

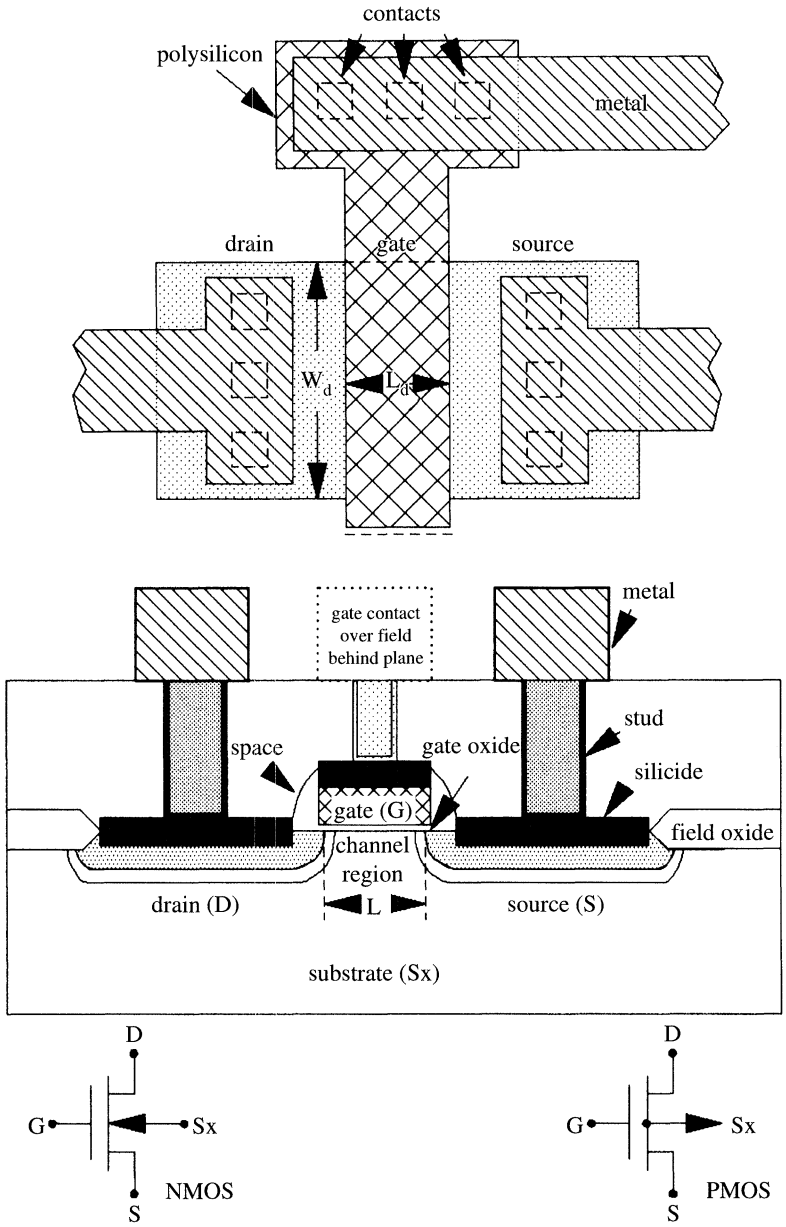

Figure 4. MOSFET top view and cross section.

(*c*) *Threshold voltage, subthreshold slope and off-current*

The threshold voltage, V_T, is defined as the gate voltage that induces a specified drain current, I_D, typically 40 nA $^*W_{eff}/L_{eff}$ for NMOS and 20 nA $^*W_{eff}/L_{eff}$ for PMOS. For gate voltages below V_T, the drain current varies exponentially with gate voltage. This region is called the subthreshold regime and the slope of $(\log I_D)$ versus V_G is called the subthreshold slope (figure 9). The drain current at zero gate voltage is referred to as the off-current I_{off}. The allowable off-current ranges from sub-pA in dynamic random access memories (DRAMs) to the nA range in logic and

Table 2. *Evolution of MOSFETs*

1965–1975 1D
NMOS/PMOS gradual channel approximation 1D bipolar transistor action spherical approximation for junction corners cylindrical approximation for junction edges
1975–1985 1D/2D
CMOS short-channel effects narrow-channel effects hot-carrier injection latch up voltage snap-back parasitic emitter edge effects
1985–1995 1/2D/3D
reverse short-channel effect reverse narrow-channel effect short/narrow-channel interactions isolation corner effects 3D memory cells 3D logic and SRAM structures vertical parasitic paths base linkage/encroachment silicon on insulator hetero-structures

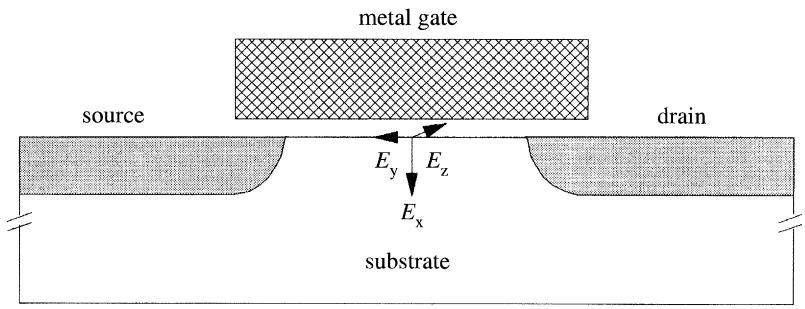

Figure 5. A 1970 MOSFET with: long and wide channel; vertical field ≫ lateral fields; gradual channel approximation; 2–3 μm deep junctions; and non-self-aligned gate and source/drain.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

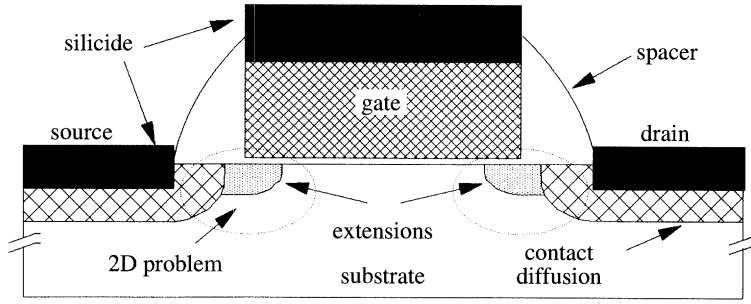

Figure 6. A 1980 MOSFET with: 0.8–1.0 μm deep, silicided junctions; 1.0 μm minimum channel length; self-aligned polysilicon gate; and lightly doped drain extensions (DD, LDD).

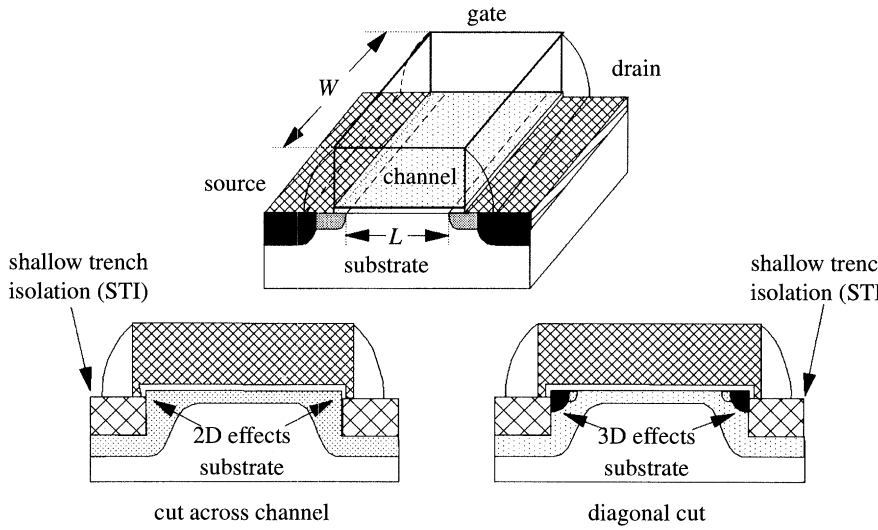

Figure 7. A 1995 MOSFET with: 0.1 μm junctions, 0.05 μm extensions; 0.2 μm channel; shallow trench isolation; and 6–8 nm gate.

static random access memories (SRAMs). I_{off} is a function of threshold voltage and subthreshold slope. Therefore, to ensure sufficiently low source to drain leakage current when the MOSFET is off, V_{T} is specified at 0.7–1.1 V for DRAMs. For logic and SRAMs, the specification of V_{T} ranges from 0.4–0.6 V and becomes a trade-off between power consumption and performance.

(*d*) *Short- and narrow-channel effects (SCE, NCE)*

For long and wide channels, V_{T} depends mainly on gate oxide thickness (t_{ox}), channel dopant profile and charges in the oxide and at the oxide–silicon interface. As channel dimensions decrease, V_{T} begins to depend on channel length and width, and on the drain voltage. The dependence of V_{T} on L_{eff} is labelled as the short-channel effect (SCE, figure 10). Lowering of V_{T} by the drain voltage is called drain-induced barrier lowering (DIBL). The narrow-channel effect (NCE) manifests itself as an increase in V_{T} as the width is decreased. SCE and NCE depend strongly on channel length, width and on the 2D–3D profiles, particularly near the drain boundary. Controlling V_{T} within tight specifications is one of the most important tasks in manufacturing.

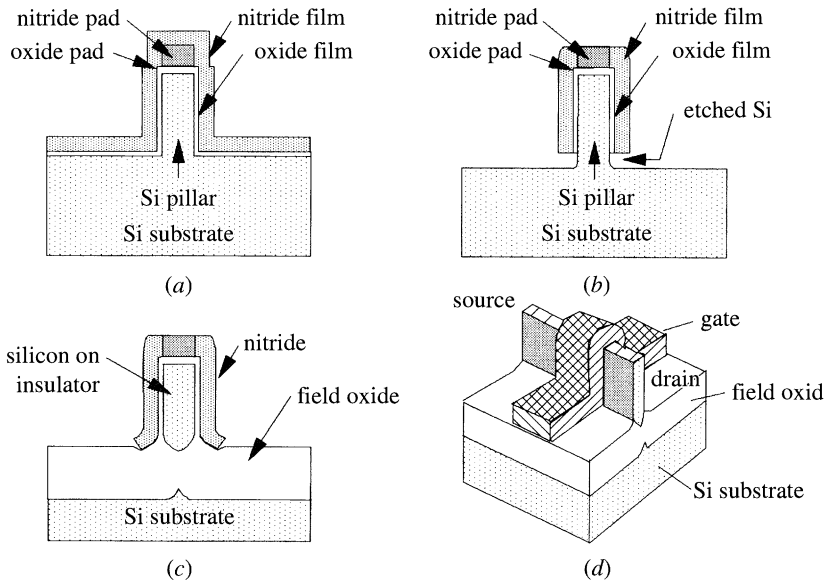

Figure 8. A year 2000 three-dimensional SOI MOSFET? (Source: D. Hisamoto *et al.* (1990).)

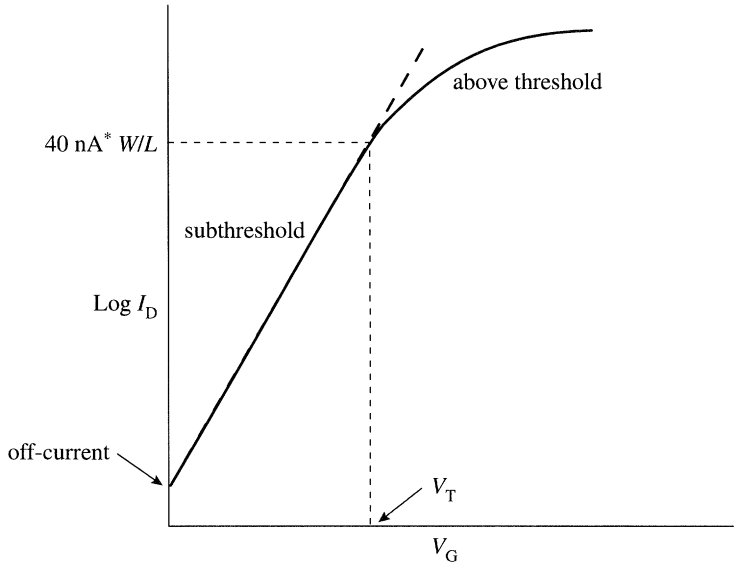

Figure 9. Definition and measurement of threshold voltage.

(*e*) *Current capability*

The maximum MOSFET current, I_{dsat}, is measured at the maximum allowable gate and drain voltages. This current determines the speed at which a capacitive load is charged or discharged. It increases as L_{eff}, t_{ox} and V_T decrease. For long and wide channels, the relation for I_{dsat} is simple. As the channel dimensions are reduced, effects such as mobility degradation by lateral fields, velocity saturation and velocity overshoot begin to affect the current-voltage characteristic, and the relation for I_{dsat} becomes more complex. The main objective of transistor engineering is to achieve the highest drain current that can be obtained at a given power supply voltage.

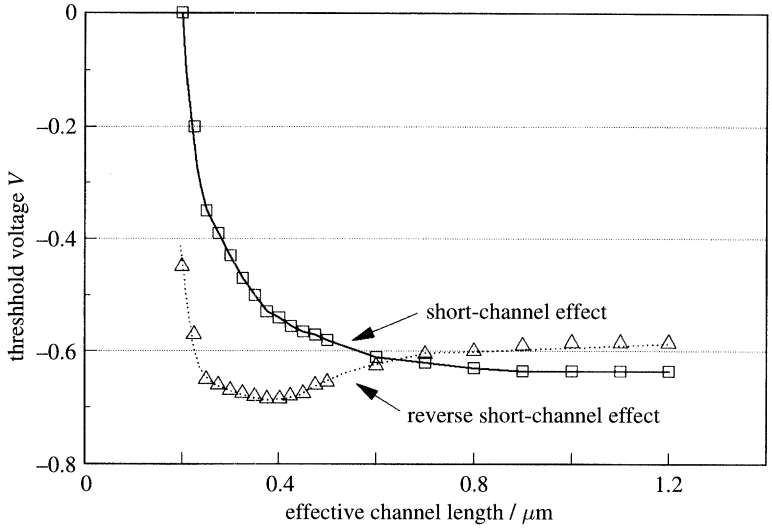

Figure 10. Short-channel effect and its reverse in PMOS.

Modulating the channel with a shallow germanium profile is an example of methods considered to enhance the transport of carriers across the channel.

(*f*) *Extrinsic resistances*

MOSFET resistances between channel boundaries and source/drain contacts are referred to as extrinsic resistances. They are composed of junction contact, series and spreading resistances at the channel boundaries. The spreading resistance is a function of the 2D–3D profile at the junction edge. Extrinsic resistances degrade the MOSFET current carrying capability since they reduce the effective gate and drain voltages and increase V_{T}.

(*g*) *Overlap capacitance*

The source and drain areas covered by the gate define a parasitic capacitance referred to as overlap capacitance, C_{ov}. While there must be some overlap to ensure current continuity, excessive overlap increases C_{ov} and reduces circuit performance, since it adds to the total load capacitance that must be charged and discharged by another MOSFET. Optimizing the 'linkage' between junction and channel is therefore critical.

(*h*) *Hot-carrier effects*

As the two-dimensional field increases at the drain boundary, carriers are accelerated and their temperature effectively increases above that of the crystal. A fraction of carriers can attain a critical energy of approximately 3.2 eV that is sufficient to overcome the silicon-oxide barrier. Part of those carriers that is directed toward the silicon surface will be injected into the gate oxide, and part of the injected carriers can be captured by energetically deep traps in the oxide or by interface states, while the rest travels to the gate. The trapped charge locally increases or decreases the magnitude of threshold voltage, depending on charge polarity. It can also degrade the carrier surface mobility. Hot-carrier effects are exacerbated by reducing L_{eff}, because then carriers arrive at the drain boundary with higher kinetic energies. A typical criterion for hot-carrier induced failures is a projected 10% change in I_{dsat} within

ten years of operation. Several techniques are implemented to reduce the 2D–3D field at the drain boundary and improve the MOSFET reliability. Among them are the reduction in power supply voltage and the formation of lightly doped drains (LDD) with gradual lateral profiles. Reducing the drain voltage by about 0.5 V, for example, compensates the effect of halving the effective channel length (Hu 1994).

(*i*) *Gate-induced drain leakage (GIDL)*

When the gate voltage is low and the drain voltage is high, the NMOSFET is off and the surface of the drain under the gate is in depletion. This condition can create a field in the overlap region high enough to cause band-to-band tunnelling in the drain and a rapid increase in gate-induced drain leakage (GIDL). For an acceptable leakage current, particularly in DRAMs, the maximum field in the overlap region should not exceed 4 MV cm^{-1}. This field increases as the dopant concentration and gradient increase at the drain surface. Careful design of the junction profile to reduce the gate field without degrading other parameters is therefore necessary. Another leakage mechanism related to the drain profile is impact ionization. GIDL and impact ionization are other reasons to implement LDD in MOSFETs.

(*j*) *Punch-through*

As the channel length is reduced, the space-charge regions of source and drain begin to merge beneath the channel, so that current can pass directly from source to drain, independent of gate voltage. For a given power supply voltage and MOSFET geometry, this condition sets the ultimate limit on L_{eff}.

(*k*) *Reverse short- and narrow-channel effects*

As the channel length is reduced, an *increase* in the magnitude of V_{T} has been observed superimposed on the 'regular' short-channel effect (figure 10). This increase, labelled as the reverse short-channel effect (RSCE), is not well understood. Several models, however, attribute RSCE to a 2D–3D redistribution of dopants in the channel. Enhanced diffusion of boron in the channel, for example, can be caused by interactions with point defects generated during polysilicon sidewall oxidation. A knowledge of the 2D–3D channel profile and dopant interactions with point defects is critical to an understanding of RSCE.

Similarly, a *decrease* in the magnitude of V_{T} as the channel width is reduced is referred to as the reverse narrow-channel effect (RNCE). So far, it has been observed only on shallow-trench isolated (STI) MOSFETs. Unlike RSCE, the reverse narrow channel effect has been well analysed and attributed to the STI 'corner field' and dopant segregation along the channel (figure 11). A step at the isolation boundaries is created during etching of oxide films, allowing the polysilicon gate to 'wrap around' the corner and induce a 2D–3D field that locally lowers the magnitude of threshold voltage. RNCE is aggravated by dopant segregation during STI processing. The threshold voltage thus varies from the STI boundary to the centre of the channel and becomes a strong function of channel width. Characterization of the 2D–3D oxide and dopant profiles around isolation regions is essential to determine processing conditions that alleviate or eliminate this problems.

(*l*) *The trade-off*

Process integration is a continuous trade-off between performance, reliability, power dissipation, 'manufacturability', cost and yield. Typically, one begins with

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

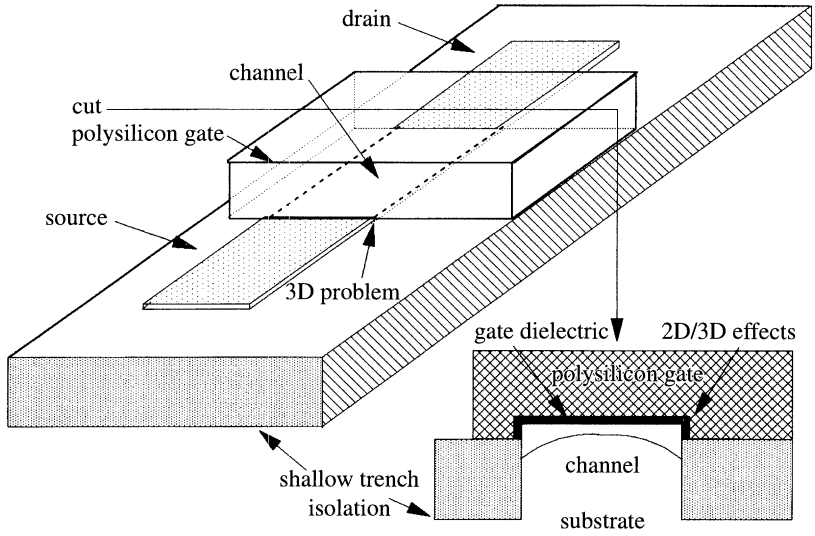

Figure 11. Two- and three-dimensional effects in shallow trench isolation.

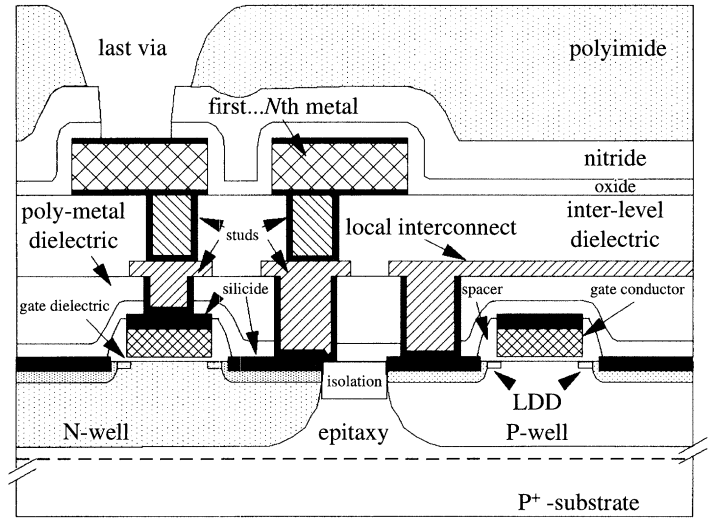

Figure 12. A typical CMOS cross section. Contact to poly typically made outside gate oxide region.

an agreed-upon standard power supply voltage. The minimum gate oxide thickness that can be reliably used with the maximum applied voltage is then chosen. The channel and drain profiles are optimized for performance, leakage and reliability. For example, reducing the LDD concentration decreases the 2D–3D field at the drain boundary, and hence decreases GIDL, impact ionization and hot-carrier injection. This is achieved, however, at the cost of performance, because of the increase in extrinsic MOSFET resistance and overlap capacitance (Bryant *et al.* 1992). Lower power consumption is a major goal for which speed must sometimes sacrificed. This is driven by the present need for battery operation, and also to reduce cost of power and heat removal systems. MOSFETs are therefore tailored to the needs and constraints

of the product to be manufactured. Such a trade-off requires intensive physical characterization and reliable simulation tools.

3. CMOS technology features

Complementarity is characterized by the presence of both n-channel and p-channel MOSFETs on the same die. A cross section of a typical CMOS structure is shown schematically in figure 12. The words 'metal', 'oxide' will continue to be used to describe, respectively, the gate conductor and gate dielectric, regardless of their composition.

(*a*) *Substrate and isolation*

The substrate is lightly doped p-type, but p^+-substrates upon which a thin p-type epitaxial layer is grown are frequently used to suppress latch-up (discussed later). A thin epitaxial layer may also be needed to improve surface roughness and reduce surface and bulk defects. N- or P-type substrates are used, depending on application. Future generations will require the reduction of metallic contaminants in the substrate to the 10^8–10^9 cm^{-3} level. While the figure shows STI for isolation, local oxidation (LOCOS) and poly-buffered LOCOS (PBL) are more frequently used for technologies with critical dimensions above 0.25 μm. The STI or LOCOS dielectric thickness ranges from 0.25 to 0.8 μm. The dopant and oxide profile at the isolation boundaries must be tailored to suppress inversion induced by a conductor passing over isolation regions, without appreciably affecting the channel profile. Encroachment into the channel can be caused by lateral oxidation and/or lateral diffusion, point defects, dopant segregation and other 2D–3D effects described above.

(*b*) *Wells*

MOSFETs are constructed on separately optimized wells, PMOS on n-well and NMOS on p-well. Three important regions are noted in each well (figure 13): the surface, subsurface and bulk. The dopant profile within about 100 nm beneath the gate oxide is tailored to achieve the specified MOSFET turn-on characteristics. The dopant profile just beneath the channel is designed to suppress short-channel effects. This region extends to about 100 nm below junctions. The purpose of retrograde well profiles in the bulk is to suppress latch-up without appreciably affecting the concentration near the surface. Typical retrograde well depths are in the order of 1 μm. MOSFET parameters depend mainly on the impurity profile near the surface. This is the most difficult region to control, model and characterize, because of unpredictable dopant/point-defect interactions and segregation effects.

(*c*) *Gate dielectric*

The gate dielectric is typically thermally grown silicon dioxide (6 nm for a 2.5 V power supply). Composite insulators, such as nitrided oxide (NO) or reoxidized nitride oxides (ONO), are also considered for deep submicron designs. The minimum equivalent gate oxide thickness is limited by reliability constraints. Oxide reliability is measured as the charge to breakdown, QBD, (or time-dependent dielectric breakdown, TDDB). While the intrinsic oxide breakdown field ranges from 1–2×10^7 V cm^{-1}, the maximum allowable oxide field is in practice constrained to 4–5×10^6 V cm^{-1} to satisfy QBD specifications.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

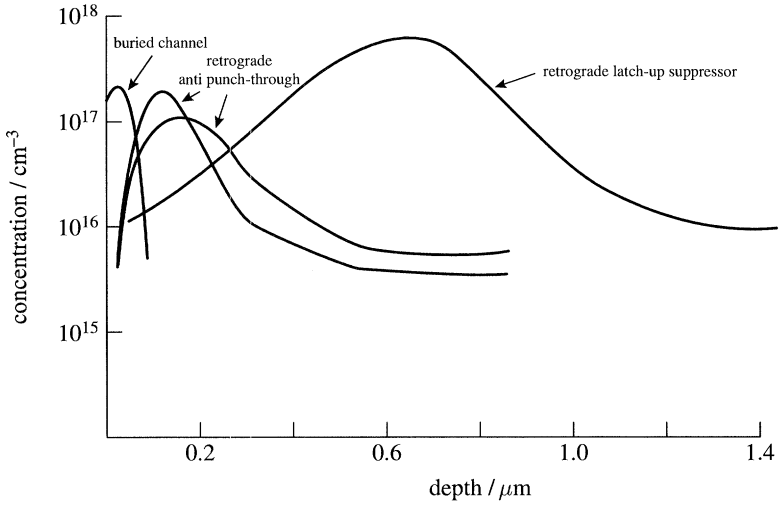

Figure 13. Retrograde n-well and surface profile.

(*d*) *Gate conductor*

The gate conductor may consist of a thin film of heavily doped polysilicon or a stack of polysilicon, silicide and barrier films, depending on application. Penetration of metals and other impurities into the gate oxide is of great concern to gate integrity and reliability. For process simplicity, n^+-polysilicon is used for both NMOS and PMOS. This necessitates implantation of a thin (50–100 nm) buried boron layer into the channel to readjust the PMOS threshold voltage. The resulting structure is referred to as buried-channel PMOS. Controlling the profile of ultra-shallow buried channels has been most difficult. The PMOS performance and GIDL can be improved by using dual workfunction gates, i.e. n^+-polysilicon for NMOS and p^+ for PMOS. The resulting PMOS structure is labelled as surface-channel PMOS. Boron in p^+-polysilicon, however, has a tendency to diffuse into the oxide and the underlying silicon, degrading the MOSFET. If polysilicon is not degenerately doped near its interface with the gate dielectric, field-induced depletion of carriers increases the equivalent gate oxide thickness and hence degrade PMOS performance. Therefore, careful annealing conditions (such as rapid-thermal annealing) and composite insulators (such as NO, ONO) may be needed to avoid degradation. Here again, simulation tools cannot predict the extent of boron penetration, and more precise analytical techniques are required to optimize processing.

(*e*) *Source and drain*

The shallow lightly doped source–drain junction (LDD) suppresses short-channel effects by decreasing the field and the extent of the lateral space-charge region at the drain boundary. LDD concentrations are in the order of 10^{18} cm^{-3}, and their depth ranges from 20 to 100 nm. Measurement of two-dimensional LDD profiles is a challenge because of the very small vertical and horizontal dimensions (50–80 nm for quarter-micron technologies), the vicinity to the surface and complex dopant interactions.

Sidewall spacers keep heavily doped source–drain junctions at a 'safe' distance from the channel. These junctions are heavily doped to reduce series and contact resistances. Large tilt-angle implants are sometimes implemented to locally increase

the well concentration at the lower lateral edge of the source and drain without appreciably affecting the surface concentration.

Ultrashallow junctions are desired to reduce the required spacer width and short-channel effects, while maintaining low contact and series resistances. Methods, such as projection gas immersion laser doping (PGILD) (Ishida *et al.* 1992), plasma doping (PLAD) (Felch *et al.* 1995) and very low-energy implants are being investigated for ultra-shallow junction formation (Sigmon & Weiner 1990). While junction silicides can further decrease incontact and series resistances, they are not compatible with ultra-shallow junctions because of metal penetration and associated leakage problems (Osburn 1990). Elevated source and drain junctions are therefore considered to form ultra-shallow junctions in silicon without encroachment into the channel, while maintaining an adequate 'buffer' for silicides and contacts. Sophisticated analytical techniques are required to optimize conditions to form such junctions.

(*f*) *Interlevel dielectrics*

A thick dielectric film is deposited over gate, source and drain to isolate the first metal from underlying structures. Three properties of the film are important: its composition, thickness and planarity. The film should have a sufficiently low dielectric constant to reduce parasitic capacitances between gate and underlying conductors. It must exhibit sufficient etch selectivity to underlying films, and be able to inhibit the migration of unwanted impurities, mainly sodium and water, to the silicon surface. The choice of insulator thickness is a trade-off between low parasitic capacitance and contact aspect ratio. Increasing the insulator thickness reduces the parasitic capacitance but the high aspect ratio (depth/width) of contact openings makes contact patterning more difficult. Also, characterization of contact openings, e.g. detection of film residues at the bottom of the contact, becomes more difficult. Contacts in a 0.25 μm technology can have an aspect ratio of 2 to 4. Higher aspect ratios are predicted for future technologies. Lower dielectric-constant materials, such as fluorinated oxides and organic materials, are being investigated to reduce parasitic capacitances without increasing the aspect ratio. Planarity is required for patterning small feature sizes because of limitations in the depth of focus of lithography tools. The most common first interlevel dielectric is borophosphosilicate glass (BPSG). It has the ability of trapping sodium ions and can reflow at about 800 °C to fill gaps and smooth the surface. Chemical-mechanical polishing is frequently used after a reflow step to ensure planarity across the full wafer. Analysis of insulator film composition and topography is essential.

(*g*) *Contacts and interconnects*

Several metal films, such as titanium and titanium nitride, ranging in thickness from 20 to 30 nm, are used as contact 'liners' to reduce contact resistance, improve adhesion to underlying and overlaying films and to act as barriers against metal penetration into junctions. Aluminium or tungsten is typically used to fill contact openings ('plugs'). Aluminium is the most widely used metal for interconnections. This metal suffers, however, from its limitation in current density because of electromigration (the dragging of metal ions by moving electrons), causing metal shorts or 'opens'. This has prompted the investigation of copper to replace aluminium at some levels, because of copper's lower resistivity (three times less than aluminium) and orders of magnitude higher immunity to electro- and stress-migration. Analysis

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

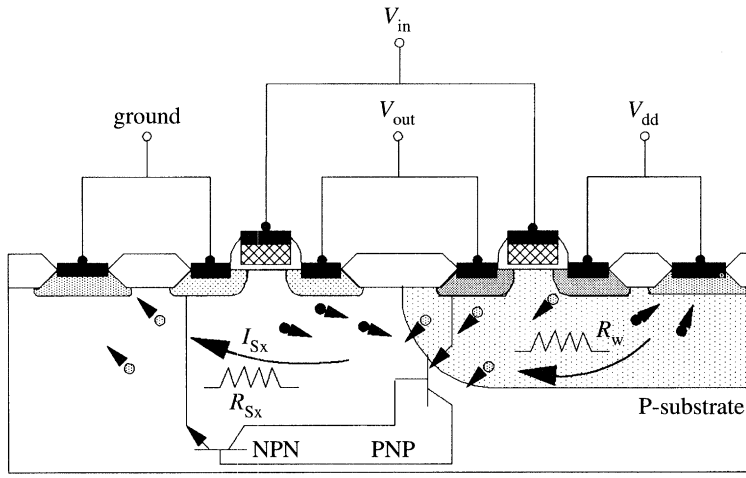

Figure 14. Latch-up in CMOS.

of the metal morphology as a function of thermal, mechanical and current stress is necessary in defining minimum metal ground rules.

(*h*) *Latch-up*

One of the major considerations in a CMOS technology is latch-up (figure 14). This is a bipolar effect in a four-layer structure, such as pnpn and npnp. Such structures do not exist in NMOS alone or PMOS alone technologies. In CMOS, they can be modelled as merged NPN and PNP bipolar transistors, with the collector of one transistor acting as the base of the other. Transients, such as a voltage 'spike' that causes junction breakdown; alpha particles that create a high density of electron-hole pairs; or inadvertent forward biasing of one of the junctions, can cause both transistors to turn on simultaneously and the structure to go from a high impedance to a low impedance mode. If the sum of the bipolar gains is larger than 1, the positive feedback may keep the structure in a sustained low-impedance mode. When this occurs between power supply and ground, the current passing is only limited by series and contact resistances. In most cases, permanent damage is caused to the structure or even the whole die. Latch-up is a three-dimensional effect that depends primarily on the junction and well dopant profiles. Latch-up immunity can be increased by reducing series resistances and bipolar current gains. This is done, for example, by implanting retrograde wells under the MOSFET junctions. Profile optimization to suppress latch-up requires careful characterization and simulation. For example, increasing the well dose improves the latch-up susceptibility, but excessive increases in dose may create defects, process delays and additional costs.

4. Three-dimensional structures

The reduction in DRAM cell size by a factor of 0.4 times every generation is only possible with the development of three-dimensional memory cell structures. A trench-capacitor cell is shown in figure 15 and a stack capacitor cell in figure 16. Both cells are used for 256 Mbit DRAM and are likely to be used for 1 Gbit and beyond (Bronner *et al.* 1995; Lee *et al.* 1995). The most important element of the cell capacitor is the node dielectric thickness and composition, where a very high dielectric constant

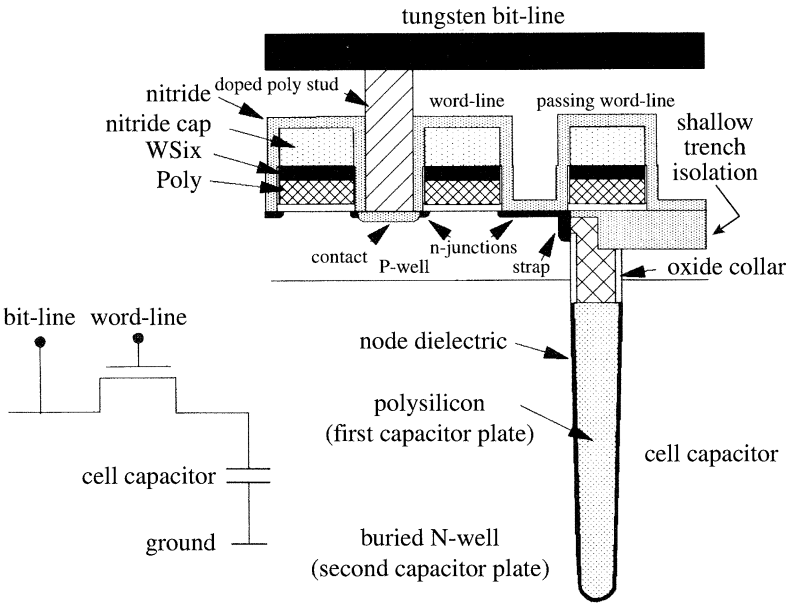

Figure 15. DRAM trench capacitor cell.

material is desired. The most commonly used dielectrics are ON and ONO. Tantalum oxide and barium strontium titanate (BST) are receiving increased attention.

To illustrate the importance of 2D–3D analysis, consider the 'buried strap' in the trench cell. This region is formed by diffusing arsenic from polysilicon inside the trench. Sufficient diffusion is necessary to ensure linkage between source–drain and polysilicon inside the trench. Excessive diffusion, however, causes the subthreshold off-current, and hence the node leakage in the '1' state to increase, reducing the cell retention time. This is a very delicate situation that requires tight control of the 3D dopant profile.

5. Silicon on insulator

Silicon on insulator (SOI, figure 17) has emerged as a high-leverage technology for a wide range of commercial and military applications (El-Kareh *et al.* 1995). Thin-film SOI has become strategic for low-power, battery-operated portable systems and large-scale integrated systems on a chip. The two most widely used techniques to form SOI wafers are SIMOX (separation by implantation of oxygen) and BESOI (bond and etch-back SOI). Several problems must be solved, however, before SOI CMOS designs enter the high-volume manufacturing stage. Among them are the availability, cost and quality of SOI wafers, the 'floating' body problem in thin-film structures, and self-heating effects caused by the low thermal conductivity of the buried oxide layer. The development of SOI-based processes depends heavily on characterization of the thin silicon bulk and surface and the buried oxide bulk and interfaces. Here again, simulation tools must be calibrated with physical analysis.

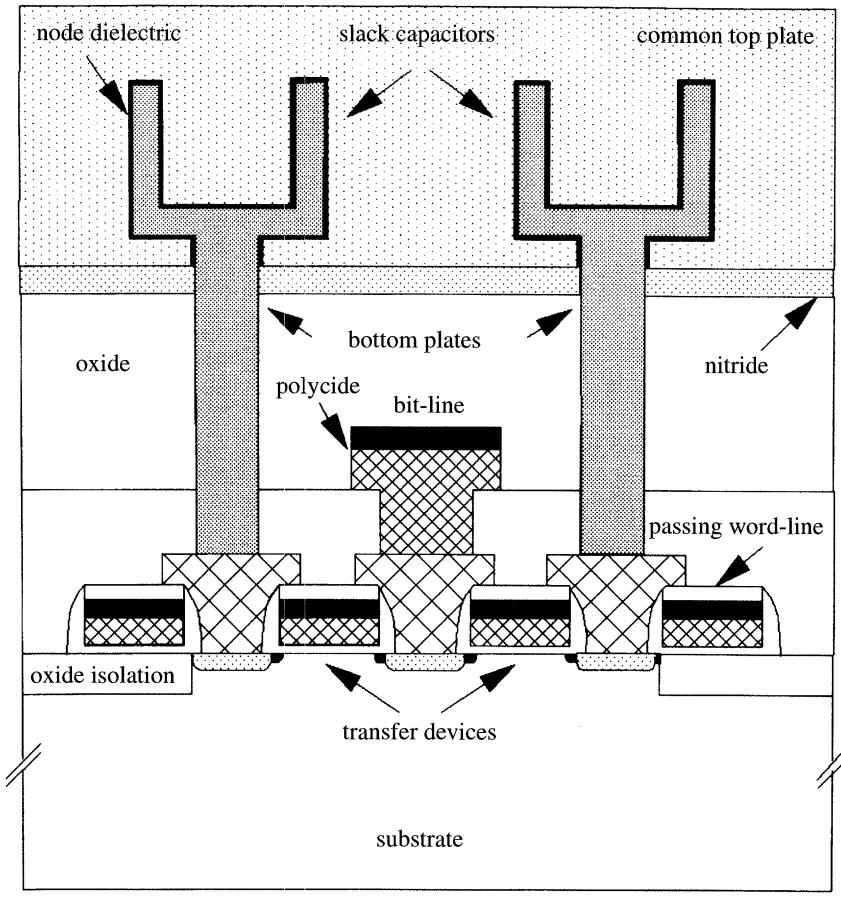

Figure 16. Three-dimensional DRAM stack capacitor cell.

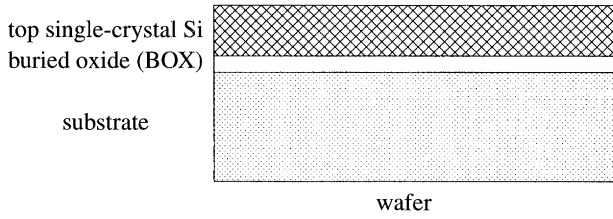

Figure 17. Typical SOI wafer, prepared by: SIMOX (separation by implantation of oxygen), high-dose oxygen implant, high temperature anneal; bond, etch-back (BESOI); zone-melted recrystallization (ZMR).

6. Defects and yield

The cost of a manufactured die is inversely proportional to product yield. Improving processing yield is therefore crucial to cost reduction. To be cost-competitive, companies must 'climb the yield curve' at a faster rate every generation (figure 1). The key to rapid yield learning is the implementation of efficient and timely in-line characterization tools to detect and analyse defect.

Yield detractors can be categorized in roughly two groups: systematic and random. Systematic yield detractors are large deviations from the desired geometries and

impurity profiles. Typical examples are film residues, line width control, over-etching, incomplete contact opening, misalignment, missing implantation or thermal cycle. As systematic errors are corrected, random defects become increasingly important yield detractors. Random defects are mostly particulates deposited onto the wafer surface or imbedded in films. The tolerable size and density of defects decreases with every generation and it becomes more difficult to detect and analyse the size and spatial distribution of those defects. More sophisticated analytical tools will therefore be required for future generations.

7. Analytical requirements

While several process simulators have been developed to reconstruct vertical and horizontal profiles in semiconductors (Selberherr 1984; Law *et al.* 1986; Pennumalli 1983; Jungling *et al.* 1985), their capabilities have lagged behind leading process development by at least one generation. There is obviously an immediate need for experimental methods to measure two- and three-dimensional profiles with adequate resolution, accuracy and repeatability for model verification and for input to device simulators. A comprehensive review of techniques to measure impurity profiles is given in Hill (1990) and Subramanyan (1992). The techniques can be grouped in roughly three categories.

1. *Staining and etching.* SEM, TEM (Sheng & Marcus 1981; Tseng & Wilkins 1987).

2. *SIMS.* Small spot-size SIMS on vertical relief (Cooke *et al.* 1989; Dowsett & Cooke 1992). 2D SIMS reconstructed from series of 1D SIMS (Goodwin-Johansson *et al.* 1992). 'Lateral SIMS' to obtain lateral dose distribution (Griegern *et al.* 1994).

3. *Electrical.* Electron-beam induced current, EBIC (Leamy 1982). Spreading resistance (Vandervost *et al.* 1992; Takigami & Tanimoto 1991). Scanning capacitive microscopy (SCM) (Williams *et al.* 1990). Overlap capacitance (MOSFETs) (Lau & Goessele 1986; Machala *et al.* 1994; Khalil *et al.* 1994) Kelvin probe combined with AFM to measure the workfunction difference between the probe and semiconductor surface and infer the dopant concentration as a function of position (Henning *et al.* 1995; Abraham *et al.* 1991).

Other techniques, such as the oxidation replica method (Hill *et al.* 1985) and lateral anodic sectioning (Kyung 1985) have also been proposed.

While considerable progress has been made in profiling techniques, the methods still suffer from one or more limitations, such as inadequate lateral resolution, need for large-area or special structures, complex sample preparation, long turnaround time, incomplete profile information, lack of reproducibility, ambiguous computer interpretation and cost. For a technique to be of value to deep submicron process development, the lateral and vertical resolution should be less than 10 nm for present designs and less than 5 nm for future technologies. The sensitivity should cover the range of 10^{15}–10^{21} cm^{-3}. The lower limit is needed to determine the location of junctions. The accuracy should be in the order of 5%. The measurement time should be less than one hour and the cost less than \$500/measurement (Rai-Choudhoury 1994).

Physical characterization of semiconductor materials will always be an integral part of process integration. The extent of measurements and requirements on analytical tools change, however, as the process moves from a research stage to full manufacturing. During research and the early stage of development, engineers would

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

be willing to sacrifice time and pay a higher cost to obtain accurate profiles and dimensions. A tool such as electron holography may be applicable to this stage. Electron holography extends the capability of transmission electron microscopy by allowing the generation of electron holograms that give information on composition and thickness of films in the angstrom range (Tonomura 1987). The present cost and time required for measurements with this tool are, however, prohibitive in a manufacturing environment.

As process development progresses, there is more focus on correlation between electrical tests and process parameters and on rapid yield learning. The need for accurate profiling and failure analysis extends throughout the development stage, with emphasis on fast turnaround and low cost of measurements.

When a product enters the manufacturing stage, the yield is typically in the 50% range, the process is 'centred', and process simulators are calibrated. Electrical tests are almost solely used to monitor the 'health of the line'. Simulations are used to relate fluctuations in electrical parameters to process variations. Analytical tools become important only when problems occur, or when dimensions are shrunk to reduce the die size or improve performance.

8. Summary and conclusions

The minimum feature size of semiconductor devices is expected to decrease by a factor of about 0.7 every three years, down to 100 nm by the end of next decade. As devices shrink, their two- and three-dimensional properties become more dominant. Process simulators are not adequate to predict these properties since they lag behind leading technologies by at least one generation. There is hence a strong need for two- and three-dimensional analytical tools, with high resolution, sensitivity, accuracy and repeatability. Requirements on the spatial resolution are more challenging with every generation, decreasing from today's 10 nm to less than 5 nm in the next decade. As development and manufacturing costs stagger, there is more focus on low-cost and fast failure analysis and in-line monitoring for rapid yield learning. The measurement time and cost per measurement are additional important considerations when choosing an analytical tool for the development and manufacturing of semiconductor products.

The preparation of this paper would not have been possible without invaluable discussions with Ed Adams, Karanam Bala, Gary Bronner, Ashwin Ghatalia, Lynne Gignac, Randy Mann, Tak Ning and Jim Ryan, all from IBM; Carl Osburn, University of North Carolina; P. Rai-Choudhury, consultant; and Jim Tompkins, SEMATECH.

References

Abraham, D. W., Williams, C., Slinkman, J. & Wickramasinghe, H. K. 1991 Lateral dopant profiling in semiconductors by force microscopy using capacitive detection. *J. Vac. Sci. Technol.* B **9**, 703–706.

Bronner, G. B. *et al.* 1995 A fully planarized 0.25 mm CMOS technology for 256 Mbit DRAM and beyond. Technical Digest, VLSI Technology Symposium, Kyoto, Japan.

Bryant, A., El-Kareh, B., Furukawa, T., Noble, W., Nowak, E. & Tonti, W. 1992 A fundamental performance limit of optimized 3.3 V subquarter micron overlapped LDD MOSFETs. *IEEE Trans. Electron Dev.* **39**, 2108.

Cooke, G., Dowsett, M. G., Hill, C., Clark, E. A., Pearson, P., Snowden, I. & Lewis, B. 1989 Two-dimensional analysis and semiconductors at dopant sensitivity using SIMS. In *Secondary ion mass spectropmetry, SIMS VII.* Chilchester: Wiley.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Dowsett, M. G. & Cooke, G. A. 1992 Two dimensional profiling using secondary ion mass spectrometry. *J. Vac. Sci. Technol.* B **10**, 353–357.

El-Kareh, B. 1994 Ultrashallow dopant film requirements for future technologies. *J. Vac. Sci. Technol.* B **12**, 172–178.

El-Kareh, B., Chen, B. & Stanley, T. 1995 Silicon on insulator, an emerging high-leverage technology. *IEEE Trans. Compon. Packag. Manufac. Technol.* A **18**, 187–194.

Felch, S. B., Sheng, T., Ganin, E., Chan, K. K., Chapek, D. L., Matyi, R. J. & Conrad, J. R. 1995 Studies of ultra-shallow p^+-n junction formation using plasma doping. In *Proc. of Ion Implantation Technology Conf. 1994*. Amsterdam: North-Holland.

Goodwin-Johansson, A. H., Ray, N., Kim, Y. & Massoud, H. Z. 1992 Reconstructed two-dimensional doping profiles from one-dimensional SIMS measurements. *J. Vac. Sci. Technol.* B **10**, 369–379.

Henning, A. K., Hochwitz, T., Slinkman, J., Never, J., Hoffman, S., Kaszuba, P. & Daghlian, C. 1995 Two dimensional surface dopant profiling in silicon using scanning Kelvin probe microscopy. *J. Appl. Phys.* **77**, 1888–1896.

Hill, C. 1990 In *Proc. Euro. Solid State Dev. Res. Conf.*, pp. 53–60. Bristol: Adam Hilger.

Hill, C., Augustus, P. D. & Ward, A. 1985 Determination of arsenic distribution in silicon by a thermal oxidation replica technique. In *Institute of Physics Conference Series 76*, § 11. Bristol: IOP.

Hisamoto, D., Kaga, T., Kawamoto, Y. & Takeda, E. 1990 A fully depleted lean-channel transistor (DELTA)—A novel vertical ultrathin SOI MOSFET. *IEEE Electron. Device Lett.* **11**, 36–38.

Hu, C. 1994 MOSFET scaling in the next decade and beyond. *Semicond. Int.* **17**, 105–112.

Hutcheson, G. D. 1995 Technology outlook: how change in chipmaking is driving inspection. KLA Yield Management Seminar, Austin, TX, USA.

Ishida, E., Kramer, K. T., Talwar, T., Sigmon, T. W., Weiner, K. W. & Lynch, W. T. 1992 *Shallow junction formation in silicon: dopant incorporation by diffusion through tungsten silicide films using gas immersion laser doping*, pp. 673–678. Pittsburgh, PA: Materials Research Society.

Jungling, W., Pichler, P., Selberherr, S., Guerrero, E. & Potzl, H. 1985 Simulation of critical IC fabrication processes using advanced physical and numerical methods. *IEEE J. Solid-State Circuits* **SC-20**, 76.

Khalil, N., Faricelli, J., Bell, D. & Selberherr, S. 1994 A novel method for extracting the two-dimensional doping profile of a sub-half micron MOSFET. 1994 Symp. VLSI Technology, Digest of Technical Papers, pp. 131.

Kyung, C. M. 1985 Two-dimensional impurity profiling near the mask edge using anodization. *Electron. Lett.* **21**, 587–588.

Lau, F. & Goessele, U. 1986 Two-dimensional phosphorus diffusion for soft drains in silicon MOS transistors. *Appl. Phys.* A **40**, 101.

Law, M., Rafferty, C. S. & Dutton, R. W. 1986 SUPREM-4: two-dimensional process modeling. Technical Report, Stanford Electronics Laboratories.

Leamy, H. J. 1982 Charge collection scanning electron microscopy. *J. Appl. Phys.* **53**, R51–R80.

Lee, K. P. 1995 A process technology for 1 Gigabit DRAM. IEDM Technical Digest, pp. 907–910.

Machala, C. F., Chern, J. H., Wise, J. L. & Yang, P. 1994 *Texas Instr. Technol. Jl* **11**, 22.

Masuda, H., Nakai, N. & Kubo, M. 1979 Characteristics and limitations of scaled-down MOSFETs due to two-dimensional field effects. *IEEE Trans. Electron Devices* **26**, 980–986.

Osburn, C. M. 1990 Formation of silicided, ultra-shallow junctions using low thermal budget processing. *J. Electron. Mat.* **19**, 67–88.

Pennumalli, B. R. 1983 A comprehensive two-dimensional process simulator program, BICEPS. *IEEE Trans. Electron Dev.* **ED-30**, 986.

Rai-Choudhoury, P. 1994 Critical review of two-dimensional profiling techniques. Report to SEMATECH.

Selberherr, S. 1984 *Analysis and simulation of semiconductor devices.* New York: Springer.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Sigmon, T. W. & Weiner, K. H. 1990 Nanosecond thermal processing. In *Proc. 2nd Int. Symp. on Process Physics and Modeling in Semiconductor Technology* (Montreal, Canada, May 1990).

Singer, P. H. 1992 Trends in CMOS development. *Semicond. Int.* **15**, 56–60.

Sheng, T. T. & Marcus, R. B. 1981 Delineation of shallow junctions in silicon by transmission electron microscopy. *J. Electrochem. Soc.* **128**, 881–884.

Subramanyan, R. 1992 Methods for the measurement of two-dimensional doping profiles. *J. Vac. Sci. Technol.* B **10**, 358–368.

Takigami, T. & Tanimoto, M. 1991 Measurements of three-dimensional impurity profile in Si using chemical etching and scanning tunneling microscopy. 1991 *Appl. Phys. Lett.* **58**, 2288–2290.

Tonomura, A. 1987 Applications of electron holography. *Rev. Mod. Phys.* **59**, 639–669.

Tseng, W. F. & Wikins, B. R. 1987 Direct observation of *N*-channel lengths. *J. Electrochem. Soc.* **134**, 1258–1260.

Vandervorst, W., Clarysse, T., Vanhellemont, J. & Romano-Rodriguez, A. 1992 Two-dimensional carrier profiling. *J. Vac. Sci. Technol.* B **10**, 449–455.

von Griegern, R., Jahnel, F., Bianco, M. & Lange-Giesseler, R. 1994 Method for the measurement of the lateral dose distribution of dopants at the implantation or diffusion mask edges. *J. Vac. Sci. Technol.* B **12**, 234–242.

Williams, C. C., Slinkman, J., Hough, W. P. & Wickramasinghe, H. K. 1990 Lateral dopant profiling on a 100 nm scale by scanning capacitance microscopy. *J. Vac. Sci. Technol.* A **8**, 895–898.

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES