Challenges and future directions for the scaling of dynamic random-access memory (DRAM)

by J. A. Mandelman
R. H. Dennard
G. B. Bronner
J. K. DeBrosse
R. Divakaruni
Y. Li
C. J. Radens

Significant challenges face DRAM scaling toward and beyond the 0.10-μm generation. Scaling techniques used in earlier generations for the array-access transistor and the storage capacitor are encountering limitations which necessitate major innovation in electrical operating mode, structure, and processing. Although a variety of options exist for advancing the technology, such as low-voltage operation, vertical MOSFETs, and novel capacitor structures, uncertainties exist about which way to proceed. This paper discusses the interrelationships among the DRAM scaling requirements and their possible solutions. The emphasis is on trench-capacitor DRAM technology.

Introduction

DRAM technology has progressed at a rapid pace since the invention of the one-transistor/one-capacitor cell (**Figure 1**) in the late 1960s [1], with an introduction of a new generation and chip density quadrupling every three years. The decade of the 1990s has seen DRAM manufacturing advance from the 4Mb to the 256Mb generation [2]. In recent years there has been a shift from a technology generation strategy (4 Mb/0.7 μm, 16 Mb/0.5 μm, etc.) to a shrink strategy (64 Mb/0.35 μm/0.25 μm/0.2 μm, etc.) with shorter development cycles [3]. The high volumes that DRAM manufacturing guarantees and the relatively predictable product roadmap have made DRAM the vehicle that drives a large part of the manufacturing infrastructure for the microelectronics industry. DRAM technology is optimized for low cost and high yield, with a particular focus on low-leakage devices and the storage capacitor.

As DRAM enters the 21st century, the course of DRAM technology development continues to be driven by the need for smaller cell sizes. To obtain a reasonable number of chips per wafer and to fit within conventional packages, DRAM chips have increased in size by about 40% per generation, while the number of bits per chip has

0018-8646/02/$5.00 © 2002 IBM

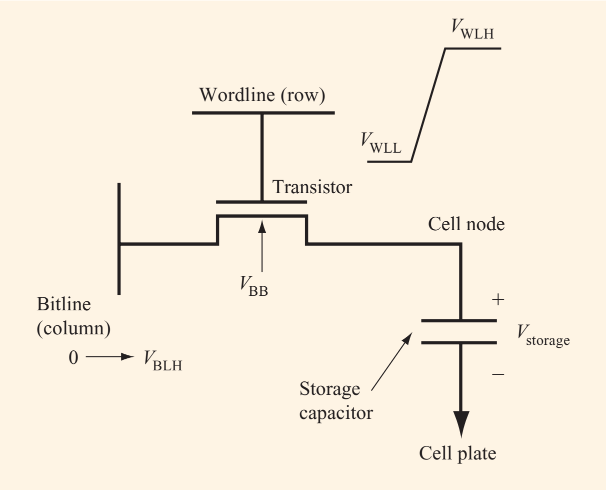

Figure 1

Schematic of a one-transistor DRAM cell [1]. The array device (transistor) is addressed by switching the wordline voltage from V_{WLL} (wordline-low) to V_{WLH} (wordline-high), enabling the bitline and the capacitor to exchange charge. In this example, a data state of either a “0” (0 V) or a “1” (V_{BLH}) is written from the bitline to the storage capacitor. V_{BB} is the electrical bias applied to the p-well.

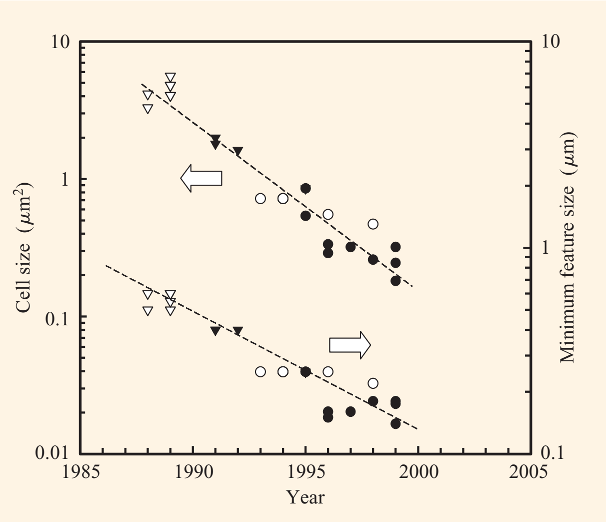

Figure 2

Progression of DRAM scaling.

increased four times in every generation. Through the 1Mb DRAM generation (prior to 1988), cell size reduction was realized primarily by reduction of the linear dimensions (i.e., minimum lithographic feature size, F). Reduction of feature size includes reduction of the wordline pitch (wordline width plus space between wordlines). The designed gate length of the array-access MOSFET is typically equal to the designed width of the wordline; therefore, decreases in wordline pitch have translated into shorter channel lengths from generation to generation. Conventional MOSFET scaling theory [4] was applied to provide shorter-channel-length MOSFETs having electrical characteristics that are satisfactory for DRAM cell-access transistors; reduction in channel length was accompanied by increased channel doping concentration and decreased gate dielectric thickness.

However, lithography scaling provides only a factor of 2 reduction in area for each linear dimension reduction of 0.7×. To achieve close to a factor of 3 reduction in cell area per generation, the remainder must come from innovations in cell structure. The 4Mb generation introduced the use of three-dimensional storage-capacitor structures [5]. From the 16Mb through the 256Mb DRAM generations (**Figure 2**), density-enhancing innovations focused on the use of techniques such as shallow-trench isolation (STI) [6], bitline contact borderless to wordline [7], and self-aligned buried strap [8]. The most aggressive 256Mb DRAM products in manufacturing in 2001 have cell sizes of approximately 0.16 μm^2, with minimum pitches of 0.28 μm and designed array MOSFET channel length of 0.14 μm (commonly referred to as the 0.14-μm technology node).

At the present point on the DRAM technology timeline, mechanisms that may limit further scaling of the channel length of MOSFETs in DRAM cells are receiving renewed attention. In order to store more charge on the capacitor, DRAM memory chips use longer channel lengths and higher voltage levels on the gate compared to the performance-oriented logic devices fabricated with equal lithography capability. Most present circuits achieve a voltage on the capacitor, V_{storage}, which is about 1.5–1.8 V less than the peak voltage applied to the gate of the memory-cell devices. Part of this voltage difference results from the high threshold voltage, V_t, in the memory-cell devices (~0.8 V) needed to prevent subthreshold conduction of charge from the capacitor to the bitline at times when the bitline is at a low voltage; body-effect and threshold-voltage tolerances add to the gate voltage required to turn on this device adequately to write the high level, V_{storage}, into the capacitor. As DRAMs are scaled to smaller dimensions, the voltage that can be applied to the memory devices will follow a path similar to that for logic devices (but delayed in time) because the DRAM devices are at a maximum field strength for gate-oxide reliability in any given generation [9]. Therefore, the stored voltage on the capacitor will shrink rapidly as the voltages are scaled down unless a better technique is found.

Another major problem which must be considered in scaling of the DRAM transistor is increased leakage due to tunneling currents in the gate insulator and in the drain–body junction. This has been shown in numerous papers [10] to be an important limit to scaling of logic transistors. It is much more critical in DRAM because of the extremely small allowable leakage [~1 femtoampere (10^{-15} A)] per device to prevent any substantial decay of the voltage stored on the capacitor. To make matters worse, the increased doping in the body of the transistor in the normal path of scaling has been shown to cause an increase in the number of transistors failing the specifications for retention of data, presumably due to some defect mechanism [11–13]. Therefore, the bounds imposed on the acceptable design space for the array-access transistor present a very serious challenge to the continued scalability of the planar MOSFET DRAM cell.

The storage capacitor is another area of focus for DRAM cell-size reduction. IBM's 256Mb DRAM chip with a minimum lithographic feature size of 0.14 μm and a cell size of 0.16 μm^2 has a storage capacitor with a surface area of approximately 5 μm^2 and a capacitance of approximately 40 fF. Through the 0.14-μm generation, methods of reducing the amount of silicon real estate occupied by the storage capacitor while maintaining sufficient capacitance have included the following: Thinning of the capacitor dielectric, use of insulating materials with a higher dielectric constant, and three-dimensional capacitor structures [14]. The ability to maintain large-surface-area capacitors in such small cells is made possible by three-dimensional capacitor structures that are built either above the silicon surface (stacked capacitors), or in the silicon substrate (trench capacitors) [15]. **Figures 3** and **4** respectively show examples of a stacked-capacitor cell (STC) [16] and a trench-capacitor cell [17] suitable for the 0.15-μm generation. It is shown in this paper that trench-capacitor DRAMs have a clear path for scaling down to design rules for less than 0.1 μm. The ability to scale stacked-capacitor cells is less clear because of challenges associated with the introduction of new dielectrics and array-device scaling problems. Since the scaling path of trench-capacitor cells appears to be more tractable than that for stacked-capacitor cells, this paper concentrates on the former.

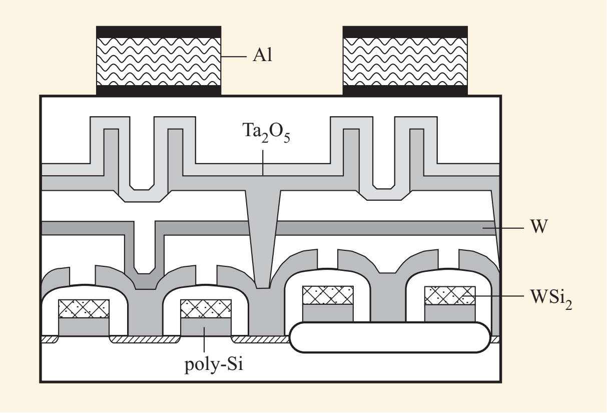

Figure 3

Schematic cross section of stacked capacitor cell suitable for 0.15 μm. Reprinted with permission from [16]; © 1994 IEEE.

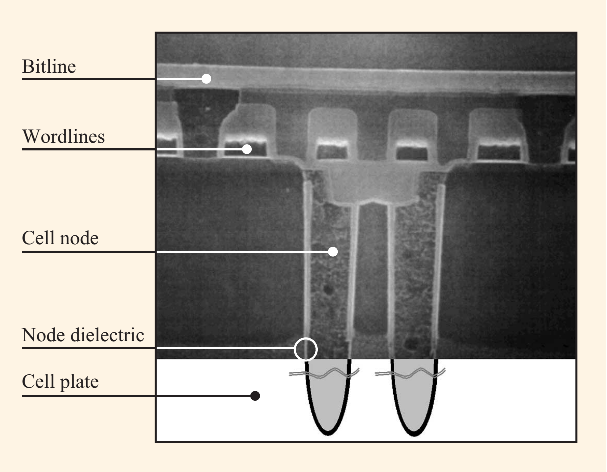

Figure 4

SEM photomicrograph of 0.25-μm trench DRAM cell suitable for scaling to 0.15 μm and below. Figure is from [17].

Trench-capacitor cells also offer the advantage of being amenable to full planarization, making trench-storage technology more favorable for integration with high-performance CMOS logic for embedded memory applications [18–20]. Integration of DRAM with high-performance CMOS logic, for embedded memory applications, is growing in importance to meet the increased data bandwidth and reduced latency requirements of speedier new generations of processors [21, 22]; however, density and performance improvements must not come at the expense of power dissipation per chip, which means that data-retention time requirements per cell remain very important.

This paper examines two important factors challenging DRAM cell-size scaling, which are driving the direction of DRAM technology development: 1) access-transistor scaling, which considers the competing requirements of threshold-voltage control, ultralow total leakage current, and MOSFET drive current sufficient for charge transfer, and 2) scaling of the storage capacitor, addressing the need to maintain adequate storage capacitance and sufficiently low series resistance.

Table 1 Array MOSFET scaling behavior, zero vs. negative wordline-low. Gate-oxide thickness, t_{ox}, is constrained by a 5-MV/cm reliability-imposed limit on gate electric field. Channel length follows the MOSFET scaling trend of being from 25 to 40× t_{ox}. For −0.5 V negative wordline-low, V_t can be reduced to about 0.3 V and still keep the transistor well turned off. For a given capacitor voltage, the negative wordline approach allows a shorter, more scaled array transistor with a lower V_{WLH} required to write a "1" into the cell. For a given value of V_{WLH}, the maximum capacitor voltage, V_{BLH}, is increased by somewhat less than 0.5 V.

Maximum device voltage, V_{WLH} (V)	*Equivalent SiO_2 t_{ox}* (nm)	*Nominal channel length, L_{eff}* (nm)	*Maximum capacitor voltage, V_{BLH}* (V)	
			V_{WLL} = 0.0	V_{WLL} = −0.5 V
3.3	6.6	250	1.80	2.28
2.5	5.0	150	1.19	1.64
1.8	3.6	100	0.63	1.07
1.5	3.0	80	0.36	0.80
1.2	2.4	60	0.21	0.65

Scaling challenges for the DRAM array transistor

A DRAM cell (Figure 1) consists of a MOSFET (also referred to as the array-access transistor or transfer device) in series with a storage capacitor. The wordline contacts the gate of the transfer device, and the bitline contacts the source/drain of the transfer device that is not connected to the storage capacitor. Data is written by turning on the transfer device by raising the wordline and writing a high or low voltage level onto the storage capacitor via the bitline. Data is stored by turning off the transfer device by lowering the wordline, trapping the voltage/charge on the storage capacitor. In industry-standard DRAM, data is conventionally read by precharging the bitline midway between the high and low levels, turning on the transfer device, and sensing the bitline voltage change (the signal voltage) caused by charge sharing between the storage capacitor and the parasitic bitline capacitance. The signal voltage is given by

$$V_{signal} = 0.5 * V_{storage} * C_{storage}/(C_{bitline} + C_{storage}),$$

where $V_{storage}$ is the voltage difference between the stored high and low levels on the storage capacitor, and $C_{bitline}$ is the parasitic capacitance of the bitline including the input capacitance of the sense amplifier.

The extent to which the actual voltage difference between the stored high and low levels on the storage capacitor, $V_{storage}$, approaches the voltage swing on the bitline (bitline-high voltage, V_{BLH}, minus bitline-low voltage, which is usually zero), is determined by the current provided by the access transistor, the value of the storage capacitor, and the amount of time allocated for the transfer of charge between the bitline and the storage capacitor. To maximize the signal, it is desired to use a value of bitline voltage swing that is as large as possible while meeting the active-power-dissipation constraints and maintaining compatibility with the chip circuitry outside the array area (the support area). As an example, in an operating DRAM, V_{signal} may be in the range of 100 to 200 mV for a $V_{storage}$ approaching 1.5 V. Furthermore, the array-access MOSFET must operate as closely as possible to an ideal switch; the lowest value of source-follower V_t for the highest drive current, while meeting the off-current objective, is desired. (As shown in Figure 1, the source-follower mode of operation occurs when charge is transferred between the bitline and the storage capacitor.) This implies a small subthreshold slope and minimal back-bias sensitivity. Although maximizing the transfer ratio $[C_{storage}/(C_{bitline} + C_{storage})]$ is also a goal, the focus of this section is on scaling the channel length of the access transistor to ever-smaller design ground rules.

Voltage-scaling issues

A scenario for scaling DRAM to smaller dimensions is shown in **Table 1**. The maximum voltage stress on the gate insulator of the DRAM access transistor occurs when either the bitline voltage or the storage capacitor voltage is zero (during writing, restoring, or reading data) and the wordline voltage is at its high level, V_{WLH}. Scaling down V_{WLH} as shown in the first column of Table 1, along the voltage-scaling path already established for logic devices, allows the effective gate-insulator thickness to be scaled down as shown for the maximum electric field of 5 MV/cm considered necessary for reliability of the gate insulator [23]. The channel length can then also be scaled down by the same amount, assuming that the depletion depth in the channel region of the turned-off device is also scaled using increased channel doping and possibly some reduction of the body bias. This reverse body bias, V_{BB} (Figure 1), is conventionally used to prevent any forward bias of the source–body junction due to circuit noise on the bitline or body, which could cause injected electrons from the source to diffuse to a capacitor node diffusion (the drain diffusion of the transistor connected to the capacitor electrode) and discharge a stored “1” level.

Referring to Figure 1, as the wordline voltage V_{WLH} is reduced, the ability to write a voltage into the cell decreases rapidly, as shown in Table 1. Two different cases are shown. In the first, the wordline for the “off” transistor is at $V_{\mathrm{WLL}} = 0$. For the second case, the “off” wordlines are kept at $V_{\mathrm{WLL}} = -0.5$ to assist further in turning off the transistor. We call this the “negative wordline-low” case. For the $t_{\mathrm{ox}} = 6.6$-nm zero wordline-low case, V_{WLH} must be about 1.5 V greater than V_{BLH} to write the full level into the cell. The required gate voltage above the sum of the source follower V_{t} and V_{BLH} is assumed to scale down with t_{ox}, thus maintaining constant inversion charge density at the end of the write “1” operation.

For the first case, the limitations on writing a “1” into the cell are shown in **Figure 5**, which plots the threshold voltage V_{t} vs. the source–body voltage for the saturated ($V_{\mathrm{DS}} = V_{\mathrm{BLH}}$) case and the linear ($V_{\mathrm{DS}} = 0$) case. To retain a stored “0” voltage on the capacitor when the wordline is at zero and the bitline is at V_{BLH} (or to retain a stored “1” when the wordline and bitline are both at zero) requires a V_{t} value of at least 0.8 V under the bias conditions identified in the figure (at the highest operating temperature) in order to keep the device current at about 10^{-15} A or less. Writing the high level (in this case 1.5 V, with a p-well bias of −0.5 V) into the capacitor causes the V_{t} to rise as shown because of the increased source–body voltage (body effect) and the reduced drain–source voltage (the reverse of drain-induced barrier lowering, or DIBL). Some amount of gate–source signal above V_{t} (about 0.3 V) is also required to keep the transistor sufficiently turned on to charge the capacitor in a reasonable time, and some allowance must be made for manufacturing process tolerances (e.g., variations in channel length, width, and STI corner effect). Thus, V_{WLH} must be about 1.5 V greater than V_{BLH} in this case to write the full level into the cell, and $V_{\mathrm{BLH}} = V_{\mathrm{WLH}} - 1.5\ \mathrm{V} = 1.5$ V for $V_{\mathrm{WLH}} = 3.0$ V.

Scaling the transistor to thinner gate oxide (t_{ox}) and reducing the maximum wordline voltage V_{WLH} will reduce the voltage V_{BLH} that can be written into the cell, as shown in Table 1. The numbers there are derived with the understanding that the minimum V_{t} value for data retention, 0.8 V, cannot be scaled, since the current at the threshold must be reduced by about eight decades and the subthreshold slope at $T = 85$°C will remain at 100 mV per decade of current change. The body effect and DIBL effect are assumed to scale down with V_{BLH}, and the overdrive and tolerances are assumed to scale down with V_{WLH} and t_{ox}. The net result is that the achievable stored capacitor voltage falls rapidly as the DRAM transistor is scaled.

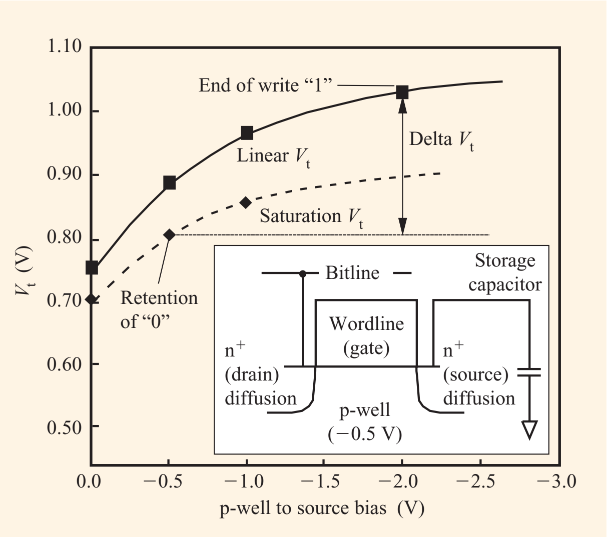

Figure 5

Illustration of the increase in array-access transistor threshold voltage between the electrical bias conditions of retaining a stored “0” and at the end of a write “1” operation. In this example, the delta in V_{t} is due solely to back-bias sensitivity and drain-induced barrier lowering (DIBL).

A somewhat better result is predicted in the negative wordline-low case, where the wordline is returned to a negative 0.5 value so that the worst-case array transistor V_{t} can be reduced to about 0.3 V and still keep the transistor well turned off. In a given generation, this lower V_{t} allows nearly 0.5 V greater voltage to be written into the capacitor, but it does require a larger wordline signal swing. More significantly, for a given capacitor voltage the negative wordline approach allows a shorter, more scaled array transistor with a lower value of V_{WLH}.

Although projections by the *National Technology Roadmap for Semiconductors* [24] call for operating voltages of CMOS logic to drop by about a factor of 0.7 to 0.8 per generation to keep power dissipation in check, DRAM designers have generally been very reluctant to consider the possibility of reducing the voltage stored on the capacitor because of the loss of signal when reading the cell and because of soft-error concerns. On the other hand, scaling principles suggest that a given design point scaled down in all dimensions and voltage should work well as far as the reduced signal level on the sense amplifier is concerned. In principle, the important noise sources are reduced with scaling, including mismatch in the sense-amplifier devices down to the point at which statistical fluctuation of impurities becomes important [10]. The eventual voltage-scaling path for DRAM depends as heavily on the capacitor as on the array

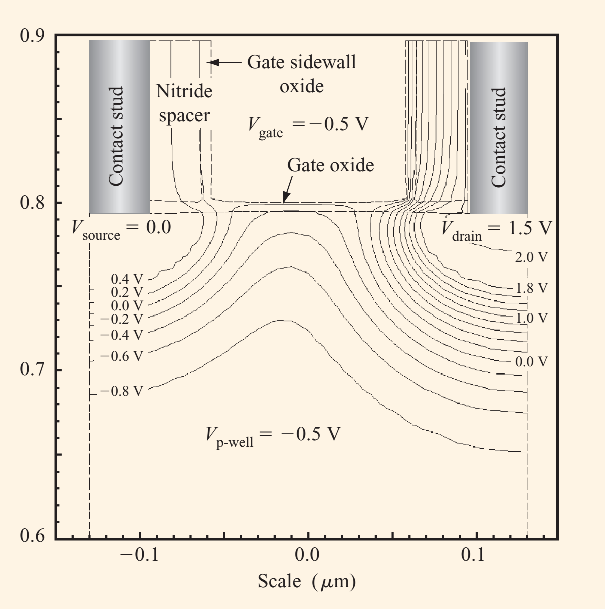

Figure 6

Modeled mid-bandgap potential contours for an exemplary DRAM MOSFET, as may be used in a stacked-capacitor cell. Note that the maximum electric field in the silicon occurs near the drain edge when the transistor is biased in the off-state. In this negative wordline-low example, the MOSFET has a physical gate-oxide thickness of 5.4 nm and a metallurgical channel length of 100 nm.

transistor. For some years, high-dielectric-constant (k) materials have been investigated and have been regarded as a future requirement for DRAM capacitors. It is quite natural that such materials could store more charge per unit area at lower voltage than today's lower-k materials.

For the sense amplifier and other support circuits to work properly at reduced voltage requires the scaling of the V_t of those devices. This will lead to the same types of off-current problems faced today in scaled logic devices, where techniques are being developed to minimize the impact on standby power. Such low-V_t devices are very achievable for DRAM embedded in a high-performance logic technology base, but until now they have not been considered affordable for industry-standard DRAM. Alternatively, different sensing circuits may be developed for lower-voltage operation [25, 26]. Ultimately, current-sensing techniques would be ideal at very low voltages to obtain the full charge from the capacitor by holding the bitline voltage nearly constant during sensing.

Another circuit design issue is posed by the negative wordline-low level, which makes it significantly more complicated to design and lay out wordline drivers in the available pitch. Interestingly, this problem could be completely obviated by a technology change to a midgap gate material (e.g., tungsten instead of n^+ polysilicon ("poly") for an n-MOSFET) that would allow identical doping profiles, electric fields, and device function, with the gate driven by the same magnitude of signal swing but with $V_{\text{WLL}} = 0$. This device would have a minimum V_t of 0.8 V, but for a given wordline swing could be scaled further (i.e., thinner t_{ox} than the n^+ poly-gated device using $V_{\text{WLL}} = 0$) because of its reduced vertical electric field. As with the negative wordline-low device, the electric field is increased in the turned-off condition relative to the grounded-wordline n^+ poly-gated case, which raises a concern about gate-induced drain leakage (GIDL) [27].

Leakage issues

All leakage-current requirements for the DRAM array transistor are much more stringent than for logic transistors. In addition to the MOSFET subthreshold off-current already discussed, several components of leakage current seen by the array storage-node diffusion are significantly affected by DRAM cell-size scaling: storage-junction-to-well leakage, array MOSFET GIDL, tunneling current in the gate insulator, and storage-capacitor dielectric leakage. The first three of these components are strongly influenced by channel-length scaling and voltage operating conditions, and the last may be affected by scaling of the storage capacitor. To ensure that adequate retention time is achieved from cell to cell across a chip, from chip to chip, and from wafer to wafer, the median value of the sum of all components of leakage current seen by the storage-node diffusion must be less than about 1 fA per cell. This ultralow value of storage-node leakage provides a guard band for the distribution of leakage, thus ensuring a sufficiently low frequency of occurrence of cells that fail to provide adequate retention time. In contrast, acceptable subthreshold off-current leakage for high-performance logic MOSFETs is typically six orders of magnitude greater than for the DRAM array device.

DRAM devices use low-dose phosphorus doping for the drain in order to achieve a low-leakage graded junction. **Figure 6** shows the simulated potential profiles in such a device biased to the turned-off condition with a negative-wordline voltage. The highest electric field in the drain–body junction occurs at the edge overlapped by the gate, where the full drain–gate voltage appears across the insulator and a depleted portion of the drain. GIDL is a leakage mechanism in which this high field can cause band-to-band tunneling in regions where the bandgap voltage is dropped across a sufficiently small distance. For either direct or trap-assisted band-to-band tunneling to be a significant contributor to leakage, the high field must occur over a distance of less than about 10 nm. According to the model of a recent reference, a field above 1 MV/cm is necessary to cause 10^{-15} A leakage current in this junction area estimated at 10^{-2} μm^2 [10]. Although the peak electric field in the silicon in this exemplary case (Figure 6) is seen to be about 1 MV/cm, only about 0.7 V is dropped over a distance of 10 nm. GIDL field reduction is also helped by the tapered oxide at the gate edge due to gate reoxidation, as shown in the figure. The rest of the junction area away from the gate edge has the usual leakage properties of a p–n junction, and the field in that region in this case is seen to be somewhat less than at the edge of the gate. As the device is scaled along the path indicated in the second column of Table 1 down to the last entry, the GIDL electric field remains fairly constant because of the compensating effects of reduced applied drain–gate voltage, V_{BLH} + 0.5, vs. thinner t_{ox}. GIDL leakage is therefore not expected to be a limitation.

However, very significantly, as seen in **Figure 7**, the higher channel doping concentration required for scaling the device results in a broadening of the cell fail-count distribution due to increased junction leakage current. It has also been reported [11–13] that increased channel doping concentration and reverse bias manifests itself as a broadening of the tail of the retention time distribution. This phenomenon is believed to be due to deep-trap-assisted tunneling. Although the origin of these randomly distributed traps has not yet been determined, point defects and/or metallic atoms have been postulated as possible causes. As seen in **Figure 8**, retention-time performance begins to degrade noticeably as channel doping rises to levels of the order of mid-10^{17} cm^{-3} [11]. This junction leakage component tends to limit the maximum doping that can be used to reduce short-channel effects and set V_t.

Because of the lower V_t, the negative wordline-low design can be accomplished with less channel doping than the grounded-wordline case. With a V_{BB} (i.e., p-well bias = −0.5 V) of the design of Figure 6, the average doping in the depletion region is about 4×10^{17} cm^{-3}. Since the band-bending at $V_G = V_t$ is fixed at $2\phi_b + |V_{BB}|$, scaling the depletion depth and maintaining the same V_t simply requires increasing the doping by the square of the scaling factor. Thus, a 1.4× reduction in dimension requires a 2× increase in doping concentration. The grounded wordline-low design requires about 2× more doping than the negative wordline-low design. Moreover, the doping must be peaked near the Si surface just below the gate oxide to avoid reducing the depletion depth (for a given V_{BB}), which would degrade the subthreshold slope. According to the reported results for the relationship between the defect leakage due to high channel doping and electric field [11–13], the reduction in capacitor voltage, which is necessary for the scaling of the array device, would possibly allow for heavier doping.

Tunneling current through the gate insulator is also a concern. With the gate of the array transistor biased to a negative value (as in Figure 6), relatively few electrons can tunnel from the gate into the weakly inverted channel, and only a portion of these will flow to the drain. The potentials are favorable for tunneling in the gate–drain overlap region, but the gate–drain insulator thickness can be locally increased relative to the t_{ox} in the channel region by the taper created in the gate-reoxidation process (also known as a gate-conductor sidewall oxidation);

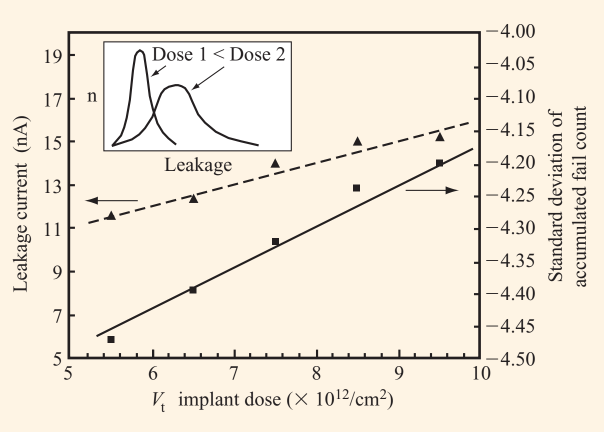

Figure 7

Increased channel doping concentration (V_t implant dose) of the DRAM array MOSFET results in a broadening of the junction leakage distribution and increased fail count. The data was obtained from the BEST [8] cell.

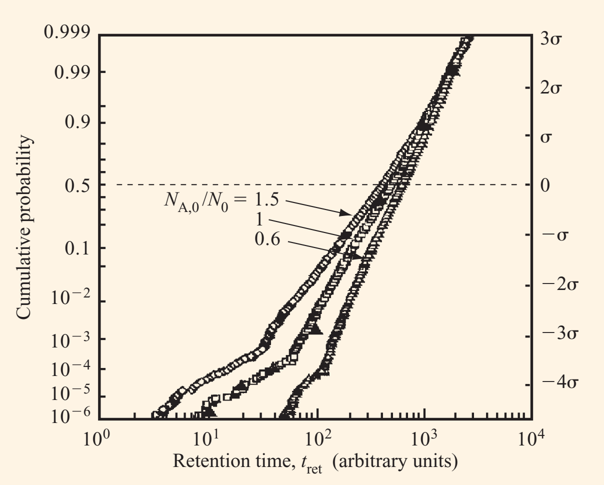

Figure 8

As the channel doping concentration, $N_{A,0}$, of the DRAM array MOSFET rises toward the mid-10^{17} cm^{-3} level, the tail of the retention-time distribution begins to degrade noticeably. $N_0 = 3 \times 10^{17}$ cm^{-3}. Reprinted with permission from [11]; © 1998 IEEE.

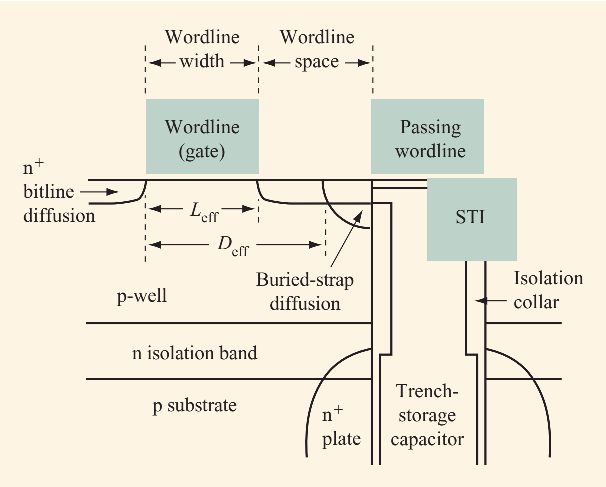

Figure 9

Schematic illustration of the BEST [8] cell, which has been the mainstay trench-capacitor DRAM cell from the 0.25-μm through the 0.14-μm generations. The presence of the buried-strap diffusion complicates the scalability of the cell.

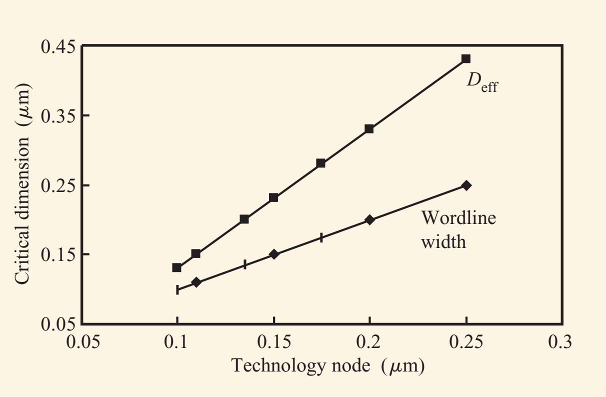

Figure 10

The distance between the buried-strap diffusion and the bitline diffusion, D_{eff}, shrinks by twice the amount of the reduction in the minimum lithographic feature size (technology node). This amplifies the DIBL sensitivity of the MOSFET as the cell is scaled. For the case shown here, the extent of the buried-strap outdiffusion from the storage-trench sidewall is 0.07 μm.

therefore, the tunneling current is greatly reduced at the edge of the gate conductor. It appears, therefore, that the most critical region is where the oxide is thinner but the channel potential is still near that of the drain. In this region, a current density of 10^{-4} A/cm^2 can be tolerated. This corresponds to a t_{ox} of about 2.5 nm for the operating voltages of interest.

DRAMs up to now have commonly used a thinner equivalent oxide for the storage capacitors in the memory cells compared to the t_{ox} used for the gate insulators. This has tended to maximize the charge stored on the capacitors, considering the lower voltage stress on them. Sustaining this trend with further scaling appears to be challenging. The leakage current requirement for the capacitor (less than 10^{-6} A/cm^2) is quite stringent because of the large area involved for trench-capacitor structures. If SiO_2 were used, the limiting thickness would be about 3 nm. The commonly used nitride–oxide composite can in principle be scaled to a somewhat thinner equivalent oxide thickness than that, perhaps 2.5 nm. This is discussed further with respect to the trench-cell technology.

On the scalability of the BEST (BuriEd-STrap) trench-capacitor cell

Buried-strap proximity

The scaling challenges for the array transistor discussed thus far are common to both stacked-capacitor and trench DRAM technologies. A proximity effect unique to the BEST [8] cell used for trench-storage DRAM from the 0.25-μm through the 0.14-μm generations degrades the V_t control of the array MOSFET. This proximity effect is due to the presence of the buried-strap diffusion, the structure of which is schematically illustrated in **Figure 9**. The self-aligned buried strap as practiced for trench-storage technology is desirable from a manufacturing cost perspective. However, it also exacerbates the DIBL effect because of both its depth and the rate at which its distance from the bitline diffusion of the access transistor varies with reduction in minimum lithographic feature size, F. Achieving a shallow buried-strap diffusion has been a challenge, since it is formed by outdiffusing dopant from the storage-trench polysilicon through an aperture on the wall of the trench [8]. The size and location of this aperture are defined by recesses of the storage-trench polysilicon, which are difficult to control relative to the minimum feature size. Furthermore, as the minimum lithographic feature size is scaled down, the proximity of the buried-strap diffusion to the bitline diffusion, D_{eff}, varies at approximately twice the rate of the reduction in the width of the wordline conductor, as illustrated in **Figure 10**.

An additional contributor to encroachment of the buried-strap diffusion upon the array-access MOSFET is overlay variation between the deep storage trench and the wordline (i.e., gate conductor) of the cell. Since the patterns for the deep-trench capacitors and the wordlines are formed from separate masking steps, requiring independent alignment, there is a statistical variation in the relative locations of these structures. Data taken from a test structure designed to intentionally introduce varying amounts of misalignment between the trench capacitor and the wordline is shown in **Figure 11**. When a high voltage is stored on the storage node, and the buried strap is close to the transfer gate, drain-induced barrier lowering due to the proximity of the buried-strap diffusion leads to an increased drop in V_t [28]. Since in the BEST cell the buried-strap diffusion is deeper than the bitline diffusion, the V_t-lowering effect is more pronounced when the array-access MOSFET is biased such that the storage-node diffusion is the drain, with the source being the bitline diffusion. The test structure used to obtain this data has a relatively long design gate length of 0.20 μm to allow the strap diffusion proximity effect to dominate and be decoupled from normal DIBL. The electrical results shown in Figure 11 suggest that the minimum useful lithographic feature size for a cell of this type is approximately 0.14 μm. At a design ground rule of 0.14 μm (corresponding to a design distance between the storage trench and the far edge of the wordline of 0.28 μm), the amount of V_t rolloff introduced by the strap proximity is approximately 200 mV. For the data shown in Figure 11, the amount of strap outdiffusion from the wall of the trench capacitor is approximately 0.08 μm, with the bottom of the strap diffusion at approximately 0.2 μm from the surface of the substrate. Although process enhancements that reduce the thermal budget and the strap outdiffusion may be introduced, control of the threshold voltage of the planar MOSFET in the BEST cell at minimum lithographic feature size less than 0.14 μm is a major challenge.

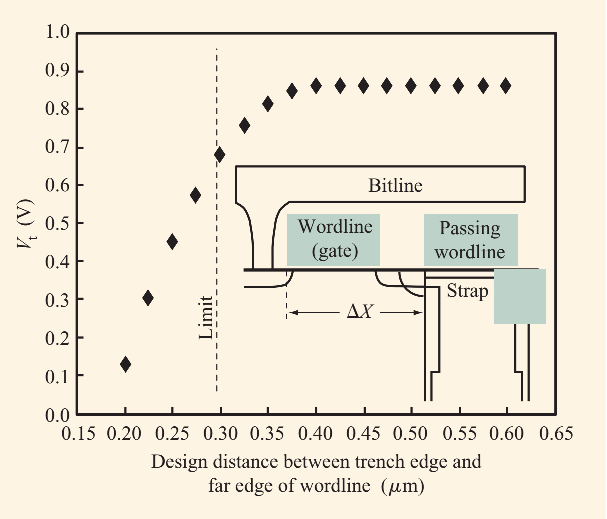

Figure 11

The proximity of the buried-strap diffusion to the array-access transistor has a strong influence on its threshold voltage. The designed channel length for this test structure is 0.20 μm, with varying amounts of storage-trench-to-far-edge-of-wordline spacing, ΔX. The relatively long channel length in this test vehicle allows decoupling of the effects of V_t lowering from buried-strap proximity and from DIBL due to drain proximity. Strap outdiffusion is approximately 0.08 μm from the edge of the deep trench.

Analysis of the manufacturing process window for the scaled BEST cell

As discussed earlier, an acceptable design point for the planar MOSFET DRAM cell must, at a minimum, simultaneously satisfy the requirements of 1) limited channel doping concentration to avoid excessive storage-node junction leakage, and 2) a subthreshold off-current of approximately 1 fA/cell. These requirements must be satisfied for all possible variations of critical physical parameters in the course of normal manufacturing process variations, and at worst-case operating conditions (i.e., temperature, voltages). The most significant physical parameters for the BEST cell [8] influencing these requirements are wordline width (i.e., gate length), alignment between the wordline and the storage trench, buried-strap outdiffusion toward the MOSFET and its depth from the top surface, and process biases and tolerances for these quantities.

The extendibility of the BEST cell [8] has been investigated by quantifying the manufacturing process design space (process window) at specific minimum lithographic feature sizes (i.e., technology nodes). In particular, the peak channel doping concentration meeting the subthreshold off-current objective for given values of wordline conductor width and proximity of buried-strap diffusion was determined. As a design guideline, the peak concentration of the channel doping concentration adjacent to the storage-node diffusion must not exceed 6×10^{17} cm^{-3}. An acceptable process window is defined as points within the range of variation of these physical parameters simultaneously falling below the 6×10^{17}-cm^{-3} channel doping limit and meeting the 1-fA off-current constraint. As indicated in **Figure 12**, variation in wordline width from the nominal value is specified as a dimensional change per edge (nm/edge), ΔGC; the total variation in wordline width would be $2 \times \Delta GC$. Encroachment of the buried-strap diffusion on the array MOSFET is characterized by the parameter α, which accounts for factors such as the extent of the strap outdiffusion from the storage-trench wall, the amount of misalignment of the wordline (GC) with respect to the storage trench, and the process bias and tolerance of the width of the storage trench.

The analyses reported here compare the relative process windows of the BEST cell at two technology nodes: 0.150 μm and 0.135 μm. Techniques such as an aggressively scaled buried-strap diffusion (60 nm outdiffusion from the wall of the storage trench and 70 nm depth from the top

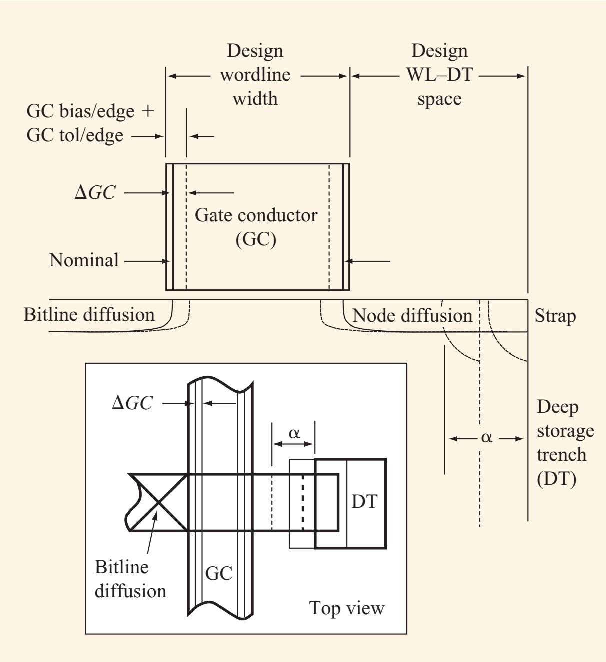

Figure 12

The manufacturing process window is characterized by parameters significantly affecting the array MOSFET off-current, ΔGC, and α. ΔGC is a measure of the deviation of the wordline (gate conductor) width from the nominal value. α is a measure of the amount of encroachment of the strap diffusion on the MOSFET, and includes factors such as the extent of the strap outdiffusion from the storage-trench sidewall, the amount of misalignment of the wordline (GC) with respect to the storage trench, and the process bias and tolerance of the width of the storage trench.

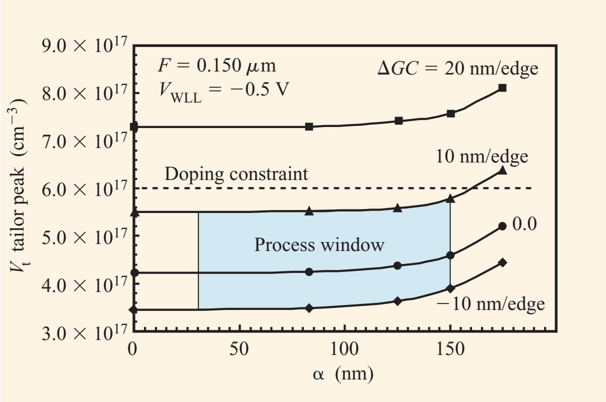

Figure 13

Manufacturing process window for the BEST [8] cell at the 0.150-μm minimum feature size. A negative wordline-low level of −0.5 V was used. Refer to Figure 12 for definition of ΔGC and α. Positive values of ΔGC and α correspond respectively to shorter gates and closer strap diffusion proximity.

surface), negative wordline-low level of −0.5 V, and equivalent gate dielectric thickness of 5.4 nm (allowing a wordline boost as high as 2.7 V without exceeding the 5-MV/cm reliability field limit) were applied to improve the process window at these small design ground rules. Process biases and tolerances representative of the state of the art were used. The devices were modeled using finite-element process [29] and device-simulation programs [30].

From the results of the analysis shown in **Figures 13** and **14**, it is apparent that the process window shrinks rapidly between the 0.150-μm and 0.135-μm nodes. At a minimum lithographic feature size of 0.135 μm, the process window is constrained by the maximum channel doping limit. Without the use of the −0.5-V negative wordline-low (i.e., customarily practiced zero wordline-low), the process window would vanish entirely at 0.135 μm. As shown in **Figure 15**, the process window may be expanded significantly by changing the wordline-low level to −0.7 V. However, the GIDL [27, 31] mechanism may impose a limit on the negative wordline-low level.

It should be noted that the planar MOSFET access transistor scales slightly better in stacked-capacitor cells than in trench-storage cells because of the absence of a relatively deep strap diffusion, whose proximity to the MOSFET is sensitive to the alignment between the gate conductor and the storage trench. STC cells eliminate the strap diffusion, since contact from the stacked capacitor is made to the top of the diffusion (Figure 3). Therefore, in STC cells the source–drain diffusions are relatively shallow. The analysis of the process window for the BEST trench-storage cell, previously discussed, considers a buried-strap diffusion depth of 70 nm with a source–drain (i.e., bitline and node diffusion) depth of 55 nm. Improved control of the buried-strap recesses in the BEST cell, or process innovations, would enable a strap diffusion depth that is not deeper than the implanted source–drain diffusions; in that case, the scalability difference between the BEST cell and a stacked-capacitor cell occupying a chip area of $8F^2$ (F^2 is an area equal to one minimum feature size long by one minimum feature size wide, $F \times F$) would be negligible.

Other thoughts on scaling the DRAM MOSFET

As discussed in previous sections, concurrently satisfying the competing requirements of ultralow off-current (~10^{-15} A for long data retention) and adequate on-current (for charge-transfer performance) is hindered by difficulties in scaling the gate-oxide thickness and the channel doping concentration in the array MOSFET. The minimum gate-oxide thickness and/or the maximum gate voltage are constrained by reliability considerations that limit the maximum allowable gate-oxide field to about 5 MV/cm. The channel doping concentration is limited by defect-enhanced deep-trap-assisted storage-node junction leakage, which degrades data retention. These limitations force the design of the array MOSFET to depart from the scalability path defined by logic-transistor technology, unless the voltage swing on the storage capacitor, V_{storage}, is also reduced. Although scaling of the storage-capacitor voltage allows for an array MOSFET that is more favorably scaled (i.e., thinner t_{ox}), it emphasizes the need for low-voltage sensing circuits, and also does not eliminate the sensitivity of junction leakage to channel doping concentration.

Negative wordline-low is effective in reducing the channel doping requirements and expanding the manufacturing process window. Although a wordline-low level as negative as −0.7 V would allow the off-current requirement to be satisfied at a significantly reduced channel-surface doping concentration, the subsurface concentration (anti-punchthrough implant) would have to remain high to contain DIBL at sub-0.10-μm channel lengths. At the same time, however, the depletion region depth must be limited such that it is properly scaled to t_{ox} and channel length, without encroaching on the highly doped punchthrough stop region. Accordingly, very sharp transitions in the channel profile as a function of depth from the gated surface would be required. Considering the large thermal budget introduced by junction anneals in DRAM processes, which have been found to be essential for reducing leakage currents, achieving such a steep channel profile may not be possible. Another constraint is that the tail of the anti-punchthrough implant must be far enough away from the depletion region associated with the storage-node diffusion to avoid increased junction leakage.

Lateral channel profile engineering (i.e., halos) has long been used to form asymmetric MOSFETs, where the channel doping is highest at one source–drain diffusion. Although this approach may be useful for limiting the channel doping at the storage-node diffusion, while meeting the V_t and sub-V_t slope requirements, for relatively long-channel array MOSFETs (i.e., longer than 0.15 μm) its implementation at scaled channel lengths may not be possible. Because of lateral channel dopant redistribution during DRAM anneal processes, it may not be possible to avoid encroachment of the halo upon the storage-node diffusion.

Geometric approaches to scaling the channel length of the array MOSFET include deviations from a purely planar channel. With the grooved gate MOSFET [32], the channel is partially contained within a groove between source–drain diffusions, with the gate conductor shielding the source end of the channel from the drain field. The step transfer device [33] also provides for drain field shielding by a partially intervening gate conductor. A third geometric variation for extending the scalability of a partially planar channel MOSFET is offered by the three-sided-gate transfer device [34]. This design provides a gate conductor which is intentionally wrapped around the sides of the narrow array MOSFET; full depletion between side gates is obtained, and penetration of the drain field toward the source is suppressed. Along the same lines as the three-sided-gate transistor, double-gate (DG-FET) [35] MOSFETs offer significantly improved scalability, but with new process-integration challenges. SOI technology has

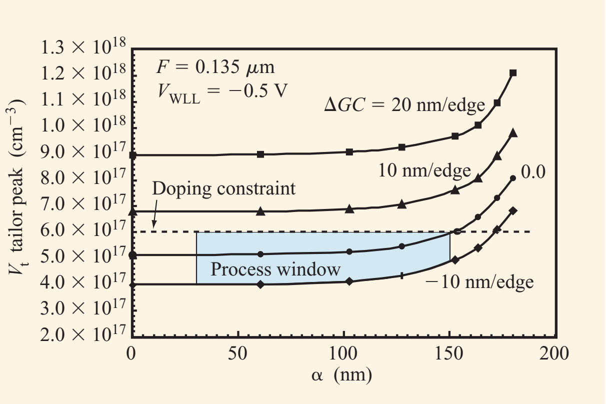

Figure 14

Manufacturing process window for the BEST [8] cell at the 0.135-μm minimum feature size. A negative wordline-low level of −0.5 V was used. The process window is only about half the size of the same cell at the 0.150-μm technology node (Figure 13). Only a maximum $\Delta GC \approx 5$ nm/edge can be tolerated.

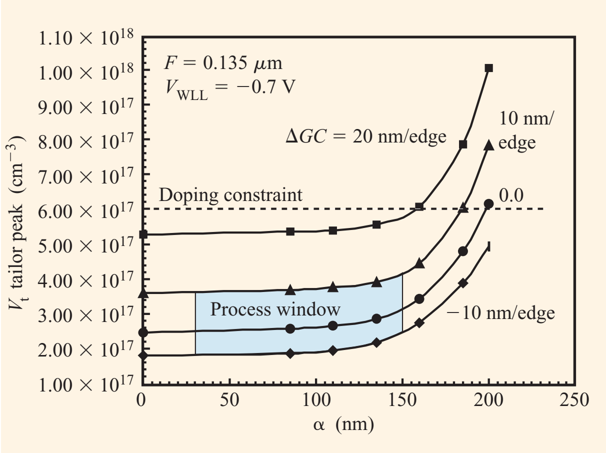

Figure 15

The process window at 0.135-μm minimum feature size is expanded by increasing the negative wordline-low level from −0.5 V to −7 V. The ΔGC range is no longer limited by the doping constraint. However, leakage contributed by GIDL [27, 31] may prevent use of wordline-low more negative than −0.7 V.

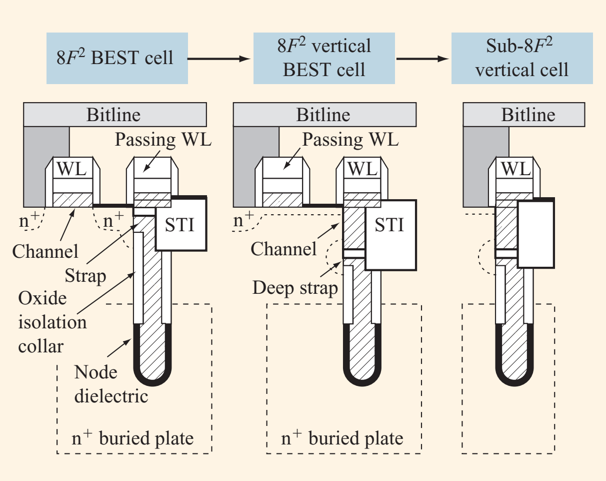

Figure 16

Evolution from the $8F^2$ planar MOSFET cell to vertical MOSFET cells, adapted with permission from [38]; © 1999 IEEE.

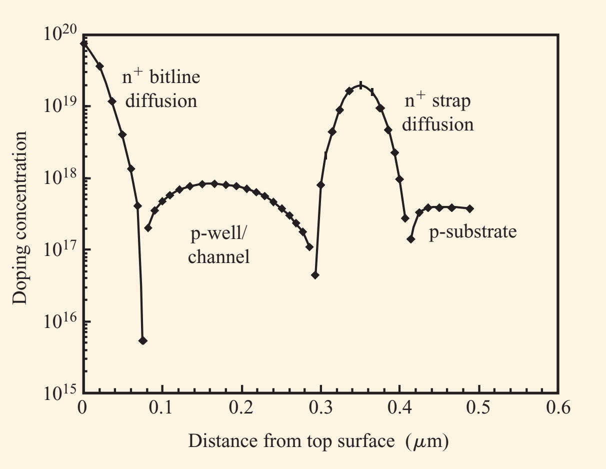

Figure 17

Modeled vertical doping profile showing that the energy of the channel doping implant can be adjusted such that its peak is sufficiently far from the strap diffusion. This helps ensure low junction leakage while still meeting the subthreshold off-current objective for long data retention. Reprinted with permission from [38]; ©1999 IEEE.

also been considered for DRAM [3] because of the benefits of reduced junction-to-body area (e.g., lower bitline junction capacitance, lower node junction leakage) and improved scalability arising from fully depleted operation. However, dynamic leakage mechanisms amplified by the parasitic bipolar transistor contained within the SOI MOSFET present a serious concern [36]. Although these geometric variations extend the scalability of a channel defined (or defined in part) by lithography, they provide only interim solutions at the cost of increased process complexity.

To sum it up, the continued scaling of the channel length of the planar MOSFET DRAM transfer device below 0.135 μm introduces new uncertainties into the picture: successful implementation of low-voltage sensing, the tradeoff between voltage scaling and increased channel doping on retention time, and extreme requirements on channel profile engineering.

In light of the discussion regarding the scalability of the planar MOSFET in a DRAM cell, there is a need to decouple the channel length of the MOSFET from the minimum lithographic feature size. It is shown next that a paradigm shift to DRAM cells using MOSFETs whose channel is oriented vertically meets this need. Although use of vertical-MOSFET DRAM cells was considered earlier [37], its adoption at the present time appears to be essential for continued reduction of cell size.

A paradigm shift—vertical-MOSFET DRAM cells

One answer to the problem of scaling the array transistor is to begin using the third dimension for the device. When a transistor is built along the walls of a trench, the channel length is decoupled from the minimum lithographic feature size and the size of the memory cell; the scaling problems for the planar access MOSFET, discussed earlier, are thus avoided. DRAM cells using trench-storage capacitors are particularly well suited for the integration of vertical transistors, since a portion of the wall of the trench above the storage capacitor is utilized for the channel, while the bitline wiring is formed above the surface of the silicon substrate. The evolution from today's BEST trench cell to a vertical-transistor trench cell is depicted in **Figure 16** [38]. A deep strap connection, including n^+ strap diffusion, is formed between the wall of the trench and the storage-capacitor polysilicon node in the trench. A second n^+ diffusion, including bitline diffusion, is formed at the top surface of the substrate, with the channel of the MOSFET on the wall of the trench between the two n^+ diffusion regions. It appears that integration of a vertical transistor with a stacked-capacitor type of cell would be more difficult to implement than for a trench-storage capacitor cell, since the stacked capacitor is formed above the surface of the silicon substrate. The bitlines would have to either run above the substrate or be buried beneath the channel of the vertical MOSFET within the substrate. The former case presents the problem of bringing the buried source–drain diffusion of the vertical MOSFET to the surface. The second option requires bitline conductors to be formed below the surface of the substrate, insulated from the substrate, and connected to the buried source–drain diffusion of the vertical MOSFET. Either stacked-capacitor case appears to involve more structural and process complexity than trench-storage vertical-MOSFET cells. Because of these complications, it is believed that trench-storage DRAM technology is the preferred approach to scaling vertical-MOSFET DRAM cells.

In vertical-MOSFET DRAM cells, the channel of the transistor is made sufficiently long to reduce threshold-voltage variations due to electrical and geometric sensitivities (i.e., DIBL) to an acceptable level. Furthermore, the relatively long channel of the vertical MOSFET allows a thicker gate dielectric that is properly scaled in proportion to the channel length, while providing reliability against wearout. Another advantage of the vertical MOSFET is that the channel doping profile may be graded such that the doping concentration in the vicinity of the buried-strap diffusion is minimized (providing reduced junction leakage) while meeting the subthreshold off-current objective needed for long data retention. As shown in **Figure 17**, the energy of the threshold implant may be adjusted to produce a peak concentration that is sufficiently far from the buried-strap diffusion.

Continued reduction in cost per bit depends upon the ability to scale the cell area more rapidly than the reduction in minimum lithographic feature size, while containing increases in process complexity. This requires that the cell be scaled below $8F^2$. The $8F^2$ vertical-MOSFET cell shown in Figure 16 utilizes a layout (**Figure 18**) in which adjacent vertical transistors are arranged back-to-back within the same region of silicon and share a common bitline diffusion. The $8F^2$ layout shown also provides good bitline noise rejection because of its folded-bitline architecture [39]. Although the $8F^2$ layout provides a space of five minimum lithographic features between pairs of storage trenches, this spacing decreases rapidly with more compact cells. A generic layout of back-to-back-storage-trench vertical-MOSFET cells is shown in **Figure 19**. For this rectangular layout, the distance between back-to-back trenches decreases from $3F$, for a $6F^2$ cell, to $1F$, for a $4F^2$ cell. One of the scalability concerns for cells using this layout is the interaction between adjacent vertical MOSFETs. Another concern is noise immunity due to the inherent open-bitline layout of sub-$8F^2$ cells.

On the scalability of back-to-back vertical-MOSFET cells

Static leakage

As the distance between back-to-back cells decreases, leakage current between strap outdiffusions, due to lowering of the potential barrier, becomes a concern; this

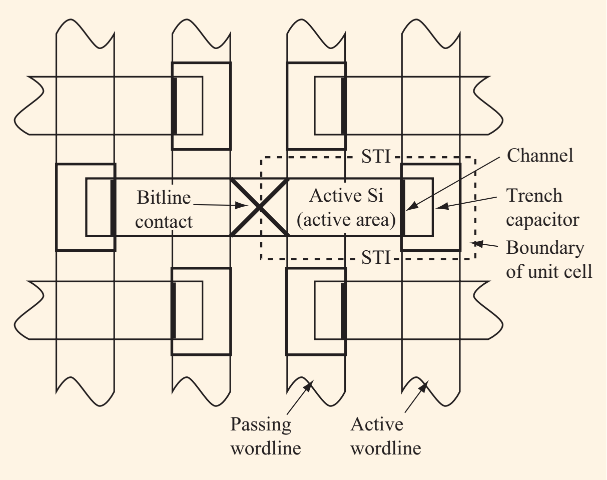

Figure 18

$8F^2$ vertical MOSFET cell layout in which the distance between back-to-back storage trenches is $5F$, adapted with permission from [38]; © 1999 IEEE.

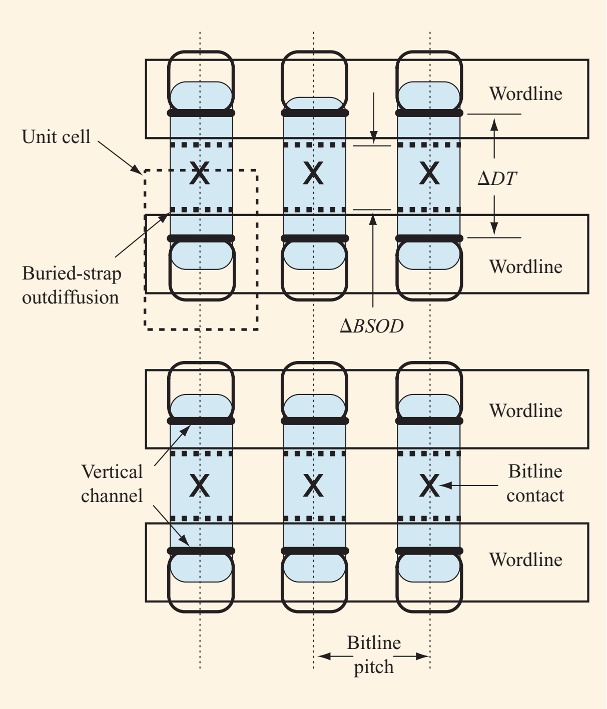

Figure 19

Generic vertical MOSFET cell top view. As cell size is reduced from $8F^2$ to $6F^2$ to $4F^2$, the distance between back-to-back storage trenches, ΔDT, decreases from $5F$ to $3F$ to $1F$. $\Delta BSOD$ is the distance between buried-strap outdiffusions of adjacent cells sharing a common bitline contact.

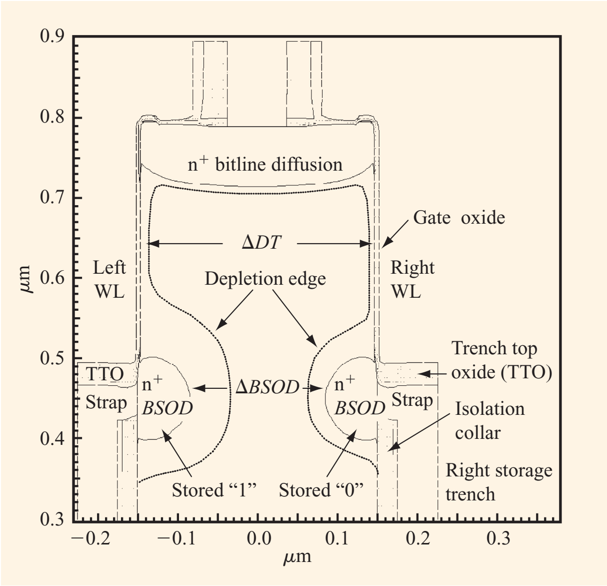

Figure 20

Modeled cross section illustrating the geometry of back-to-back storage trenches and approximate depletion region edge. A high level ("1") and a low level ("0") are stored on the left-hand and right-hand capacitors, respectively.

is a manifestation of the drain-induced barrier-lowering mechanism (DIBL), due to penetration of the electric field, which is well known for MOSFETs [40]. The modeled geometry and approximate location of depletion region edges of back-to-back vertical-MOSFET cells are shown in **Figure 20**. To first order, the extent of the barrier lowering is a function of the p-well doping concentration between n$^+$ strap diffusions, and the distance between metallurgical junctions; the highest p-well concentration and smallest strap outdiffusion are desired. This static leakage mechanism results in an adjacent stored high level (i.e. "1") and low level (i.e. "0") leaking toward each other, which results in degradation of signal margin.

Since the maximum p-well doping concentration is constrained by junction leakage considerations, scalability of this cell depends on minimizing the extent of the buried-strap outdiffusion.

Dynamic leakage

Also of concern in scaling a cell of this type is a dynamic leakage mechanism for a stored "1" when the bitline and the wordline of the adjacent cell are cycled in the course of data read, write, and refresh operations. As the cell is cycled, the distribution of majority carriers (i.e., holes) in the p-well region between the two opposed vertical gates is modulated by the time-varying electric field. Majority carriers must be able to flow freely between the portion of the p-well between the gates and the region below the strap diffusions to maintain charge equilibration in the well. The undepleted region between two back-to-back buried-strap outdiffusions narrows as the spacing between storage trenches in two back-to-back cells is reduced, thus impeding the flow of holes and pumping the voltage on the p-well between the gates as the wordline is cycled.

Modulation of the voltage on the p-well and the buried-strap diffusion (as the data state is changed) of the adjacent cell results in a loss of charge from the stored "1" due to both transient subthreshold leakage and transient exchange of electron charge between the two adjacent strap diffusions. Finite-element device simulation [30] of the cell has determined that the worst-case data pattern for a loss of a stored "1" occurs when the adjacent cell is repeatedly cycled between write "1" and write "0"; most of the loss occurs when a "0" is written over a cell with a stored "1." It may seem unlikely that a DRAM cell would be exercised repeatedly with a write "1"–write "0" pattern; nevertheless, such a pattern is possible, and data integrity must be ensured. Although the loss of charge

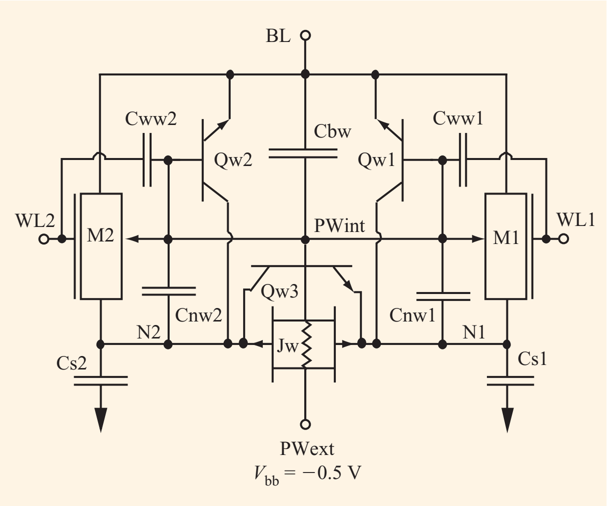

Figure 21

Equivalent circuit of back-to-back vertical MOSFET cells. M1, Cs1 and M2, Cs2 are respectively the access transistor and the storage capacitor of the right-hand and left-hand cells. The equivalent circuit contains many parasitic elements which account for the signal-loss mechanisms. Jw is a parasitic JFET whose channel represents the undepleted region between storage-node diffusions (i.e., buried-strap outdiffusions) N1 and N2. Qw1, Qw2, and Qw3 are parasitic npn bipolar transistors which may conduct when the well potential between access MOSFETs (Pwint) is allowed to bounce due to coupling from a cycling wordline and storage node. A high voltage (i.e., "1") stored on Cs2 may degrade primarily because of conduction of Qw3.

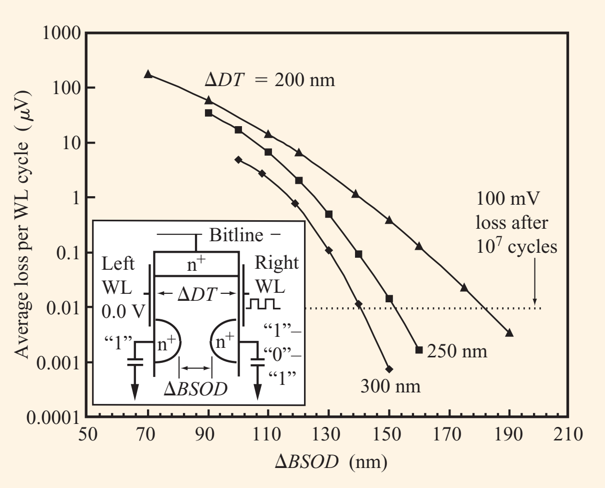

Figure 22

Modeled average loss of a stored "1" per cycle due to repeated write "1" – write "0" pattern on adjacent cell.

from a stored "1" may be less than a tenth of a microvolt per cycle, inability to detect the "1" may occur since 10^6 to 10^7 wordline cycles are possible before data is refreshed. The interaction between cells is more easily understood with the assistance of the equivalent-circuit model shown in **Figure 21**. **Figure 22** shows the average loss per cycle of a stored "1" due repeated cycling between write "1" and write "0" on the adjacent cell. The rate of loss of a stored "1" has been calculated with a full 1.5 V on the storage capacitor as a function of spacing between back-to-back buried-strap outdiffusions, $\Delta BSOD$, with the spacing between storage trenches, ΔDT, as a parameter. The minimum acceptable end of process ΔDT, arbitrarily considering that the maximum acceptable loss of a stored "1" is 100 mV, is indicated after 10^7 wordline cycles. The rate of loss decreases slightly as the strength of the stored "1" is reduced because of contraction of the depletion region and expansion of the undepleted width between strap diffusions. **Figure 23** shows the minimum usable feature size (F), as constrained by both static and dynamic leakage mechanisms, as a function of normalized cell area (i.e., number of square minimum features). A typical buried-strap outdiffusion of 50 nm from the trench sidewall is considered. It is noteworthy that the dynamic leakage mechanism sets the constraint for the minimum feature size. These results based on conservative assumptions support scaling of the $6F^2$ cell to ground rules smaller than 0.09 μm, for a cell size smaller than 0.05 μm^2.

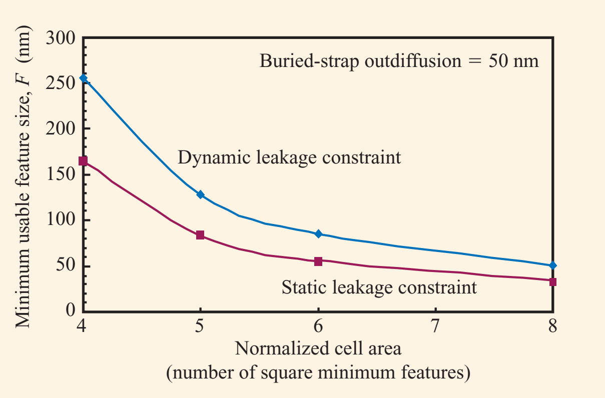

Figure 23

Tradeoff between the minimum allowable feature size (F) and the normalized cell area, expressed as the number of square minimum features, for the layout shown in Figure 19. As the normalized area of the cell is reduced, the distance between back-to-back storage trenches decreases and the minimum usable feature size must be increased. Dynamic leakage is the limiting mechanism. A $6F^2$ layout is acceptable at $F \approx 90$ nm.

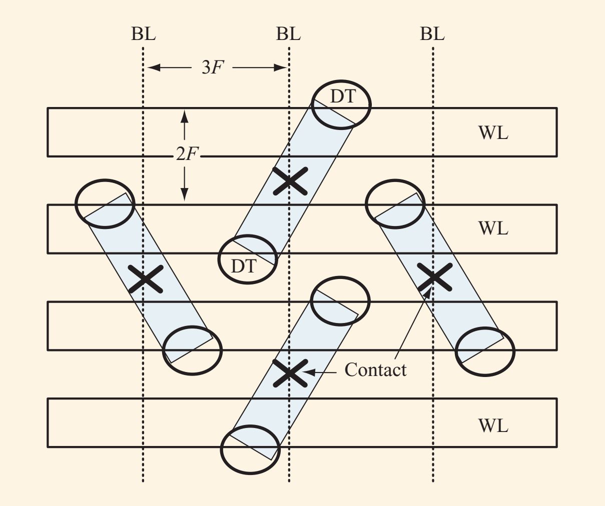

Figure 24

Herringbone active area (AA) pattern for a $6F^2$ vertical MOSFET cell, adapted with permission from [41]; © 2000 IEEE. The bitline pitch is increased to $3F$, while nearly $3F$ spacing is maintained between trenches (DT). The $3F$ bitline pitch facilitates the layout of the sense amplifiers for this open-bitline architecture.

As indicated by Figure 23, for a given buried-strap outdiffusion, the scalability of the cell is limited by the minimum allowable spacing between back-to-back trenches. Alternative layouts can be used to increase the bitline pitch while maintaining desired cell area and trench-to-trench spacing. The $6F^2$ cell shown in **Figure 24**

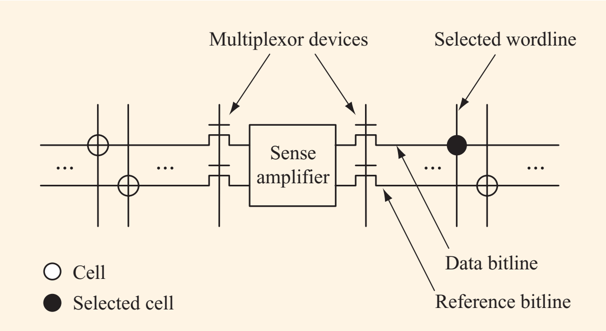

Figure 25

Folded-bitline array architecture.

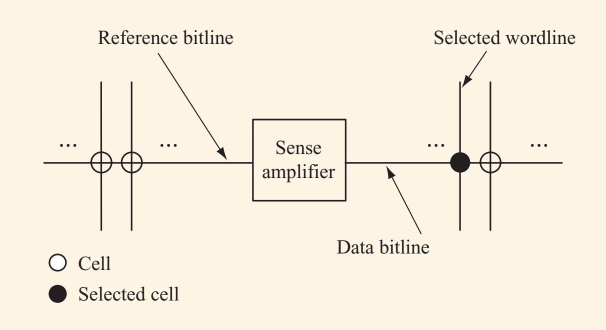

Figure 26

Open-bitline array architecture.

[41] uses a herringbone active-area (AA) pattern to increase the bitline pitch to $3F$ (from $2F$ in the rectangular layout shown in Figure 19), while maintaining close to $3F$ spacing between trenches. The increased bitline pitch is important with the $6F^2$ cell because it facilitates the layout of the sense amplifiers. While $8F^2$ layouts having a folded-bitline architecture pose no problem for sense-amplifier layout with a single level of bitline wiring, the inherent open-bitline architecture of $6F^2$ cells requires two levels of bitline wiring to achieve folded-bitline operation.

Array architecture considerations for sub-$8F^2$ cells

Continued reduction in cost per bit of DRAMs depends upon scaling of cell sizes faster than the scaling of lithography features alone would provide. Ever-shrinking cell size also demands fundamental changes in the array architecture. Folded-bitline array architecture (**Figure 25**) has been used universally within the DRAM industry since the 1Mb era. In this architecture, the cell contains one bitline and two wordlines. The "active wordline" forms the gate of the transfer device, whereas the "passing wordline" simply passes over the cell. The cells are arranged so that the passing wordline of one cell becomes the active wordline of the adjacent cell, and vice versa. Thus, when a wordline is selected, signal charge is released onto every other bitline. Each sense amplifier serves two adjacent bitlines, allowing each sense amplifier to sense the "data bitline" using the adjacent bitline as a reference. The adjacent nature of the data and reference bitlines provides excellent matching and noise rejection. By including multiplexor devices on either side of the sense amplifier, each sense amplifier can be shared between two bitlines from the left array and two bitlines from the right array. Thus, each sense amplifier serves a total of four bitlines.

The folded-bitline architecture requires a minimum cell size of $8F^2$, where F is the minimum lithographic feature size of the technology. The minimum cell size equals one minimum-bitline pitch (1 line + 1 space = $2F$) times two minimum-wordline pitches (2 lines + 2 spaces = $4F$), or $8F^2$. Until recently, this limit had not been reached, as other structures within the cell (bitline contact, transfer device, node diffusion, storage capacitor, and isolation) required more than $8F^2$ and therefore limited the cell size. However, since DRAM cell area historically scales more quickly (0.33× per generation) than lithographic improvements alone provide ($0.70^2 = 0.49\times$ per generation), the cell size measured in F^2 must decrease over time. The cell size reached the $8F^2$ limit at the 0.175-μm generation, with these technologies being qualified in late 1999 or early 2000. For the cell area to continue to scale at the historic rate, the $8F^2$ limit must be overcome. The vertical cell described in this paper allows the other structures within the cell to fit into less than $8F^2$, but does not address the problem of providing the wiring required for folded-bitline operation. DRAM designers around the world have been aware of this approaching limit for some time. Many solutions have been proposed; however, no consensus on a preferred solution has been established.

The most obvious solution is the open-bitline architecture seen in **Figure 26**, which was used universally within the industry prior to the introduction of the folded-bitline architecture. In this architecture, a cell contains one bitline and one wordline, and the sense amplifier senses the data bitline using a reference bitline from the adjacent, inactive array. The minimum cell size is $4F^2$, although other constraints within the cell (i.e., minimum trench-to-trench spacing) will limit the cell size to something greater than $4F^2$ for some time. However, the open-bitline architecture has several drawbacks. Since the data and reference bitline are located in adjacent arrays, the matching and noise rejection characteristics are not as favorable as those of the folded-bitline architecture. Further, since each sense amplifier serves only two bitlines as opposed to four bitlines for the folded-bitline architecture, the chip will require twice as many sense amplifiers, assuming the same number of bits per bitline. This is a significant drawback, since sense amplifiers occupy approximately 10% of the total chip area in a modern DRAM.

After evaluating the open-bitline option and several other sub-8F^2 DRAM array architectures [42–44], IBM has focused its efforts on the vertically twisted bitline array architecture seen in **Figure 27** [45–48]. In this architecture, the reference bitline is located directly above the data bitline on the subsequent level of metal. For this reason, this architecture requires an additional level of metal for the second level of bitlines. The cells are attached to the lower of the two bitline levels. The cell contains two bitlines and one wordline, yet the minimum cell size which can be wired is 4F^2, since the two bitlines are on different levels. At intervals along the length of the bitline, the two bitlines exchange levels, or "twist." Each bitline is on each level for equal lengths, so that the two bitlines are matched. As in the folded-bitline architecture, the data and reference bitline are within the same array, leading to good matching and noise rejection. Also as in the folded-bitline architecture, the sense amplifier can be shared between two arrays, allowing each sense amplifier to serve four bitlines. Since the twists involve only wiring and contacts, and no devices, the underlying array need not be interrupted, and thus there is no area penalty for the twists. This architecture is particularly well suited to IBM's trench-DRAM technology, since the capacitor structure of a stacked-capacitor DRAM technology would likely interfere with the second level of bitlines and the twist contacts.

A very significant advantage of this architecture is the cancellation of bitline coupling noise, as shown in **Figure 28**. When the twists are staggered relative to those of the adjacent bitline pairs, noise from adjacent bitlines couples equally into the data and reference bitlines, eliminating a significant differential noise source. This is accomplished by manipulating the matrix of inter-pair coupling capacitance such that the capacitance from any bitline "BLi" to any other bitline "BLj" exactly matches the capacitance from "i" to "j's complement," BLj(bar). Therefore, any change in the voltage on BLi couples equally into BLj and BLj(bar), creating no differential noise on pair BLj. This method reduces bitline coupling noise during both the initial signal development and the subsequent signal amplification. Similar methods have been used by IBM and others to reduce bitline coupling noise in the folded-bitline architecture [49].

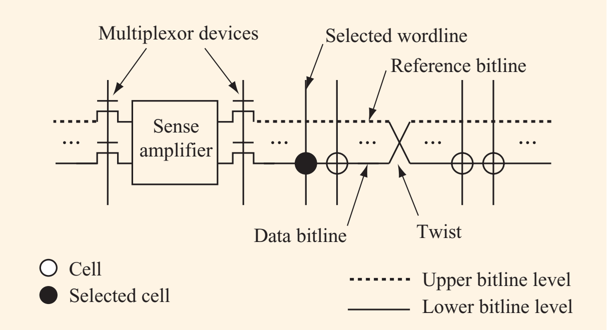

Figure 27

Vertically twisted bitline array architecture.

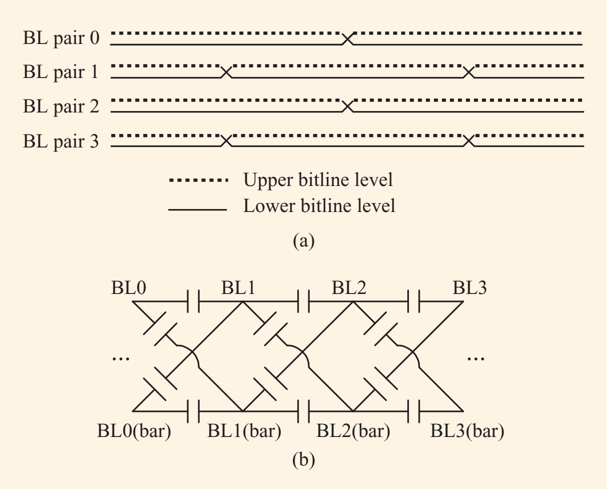

Figure 28

Cancellation of bitline coupling noise: (a) Staggered twist arrangement; (b) resulting matrix of interpair coupling capacitance (all capacitors of equal value).

Storage-capacitor scaling challenges

Storage-capacitor requirements

In addition to reducing the area occupied by the DRAM access transistor, cell-size scaling depends on the ability to scale the portion of the cell layout area allocated to the storage capacitor. The voltage swing on the storage capacitor, $V_{storage}$, has scaled more slowly than the reduction in minimum lithographic feature size because of concerns about providing sufficient overdrive to set the sense amplifiers quickly. Furthermore, the signal induced on the bitline, V_{signal}, has been constrained to be at least 100 mV by concerns about reliably setting the sense amplifiers in the presence of various sources of noise (i.e., sense-amplifier V_t mismatch, coupling of switching disturbances, leakage, transient radiation-induced charge, etc.). Since the 4Mb generation, bitline capacitance has generally remained within a range from about 150 fF to 350 fF, since the reduction in capacitance per bit resulting from scaling of physical dimensions of the bitline conductor and the cell width has been offset by increases in the number of bits per bitline, with increasing bits per chip.

The storage capacitance must be sufficient to ensure the objective of minimum signal induced on the bitline, V_{signal}; this is dependent on the product of the transfer ratio and the voltage swing on the capacitor, $V_{storage}$. As a result of the constraints discussed in the previous paragraph, the required storage capacitance is expected to remain in the range from 30 to 40 fF. Therefore, the first challenge faced by all DRAM manufacturers is fabricating a low-leakage storage capacitor with adequate capacitance in an ever-decreasing cell area.

Planar storage-capacitor structures were employed through the 1Mb DRAM generation. With the arrival of the 4Mb DRAM, adequate cell capacitance could no longer be obtained from simple planar capacitors, marking the introduction of trench- and stacked-capacitor structures. In the mid-1980s, IBM, Texas Instruments, and Toshiba introduced the trench capacitor; stacked-capacitor designs were shown by other DRAM manufacturers at about the same time. These three-dimensional trench-capacitor and stacked-capacitor structures were developed to maintain the capacitance of the array cell relative to the bitline capacitance as the cell size and bitline wiring spacing were reduced with successive DRAM generations. In IBM DRAM products, deep-trench storage has provided a remarkable platform for the continuous scaling of the charge-storage element. The trench structure has served to decouple the effective surface area of the capacitor, and hence its capacitance, from the area of the array cell, while maintaining 35 fF–45 fF per cell.

The total capacitor surface area available in a stacked design is considerably less than that in a trench capacitor. This comes about because the height of a stacked-capacitor cylinder as shown in Figure 3 is limited to 1–1.5 μm. Anything taller than this causes problems with mechanical stability. Additionally, it becomes difficult to wire over the topography of such a tall capacitor. This leads to a need to introduce capacitor dielectrics with higher dielectric constants (more capacitance per unit area) than the NO (nitride–oxide) dielectric commonly used by DRAM manufacturers through the 0.15-μm generation, for both stacked and trench designs. Stacked-capacitor manufacturers have been introducing a series of new structures and materials for smaller-ground-rule technologies. Conventional NO dielectric has been adequate down to 0.15 μm. After that, Ta_2O_5 is being introduced for the 0.12-μm generation. For the 0.1-μm generation, a new material with yet higher dielectric constant (relative dielectric constant >20) will be needed. Today no material has yet been shown to be adequate for this generation, but most companies have been researching barium strontium titanate (BSTO) as the most likely candidate. Although the scaling path for trench capacitors appears to be more certain than that for stacked capacitors, there are many challenges that must be addressed.

In addition to minimum capacitance requirements, the series resistance of the storage capacitor must be contained such that it does not degrade the transfer of charge between the bitline and the capacitor. The current through the array MOSFET is influenced by the total series resistance of the cell: channel on-resistance of the MOSFET, source–drain resistance, resistance of the bitline contact and of the connection between the MOSFET and the capacitor (the strap in the case of a trench capacitor), and the series resistance of the capacitor itself. As the minimum lithographic feature size is scaled in a trench-capacitor cell, the series resistance of the conductor in the storage trench becomes the dominant contributor to the total resistance. For stacked-capacitor cells, the resistance of the interconnection between the node diffusion of the MOSFET and the overlying capacitor is a scaling concern. To avoid degrading the charge-transfer performance, the series resistance contributed by the capacitor must not generally exceed about 50 kΩ.

Each physical component of the trench capacitor introduces potential limits to its continued aggressive scaling. Areas of concern include the processing (i.e., etching and filling) of the increasingly high aspect ratio of the trench, thinning the capacitor dielectric for higher capacitance while limiting leakage, reducing capacitance loss due to formation of depletion regions on the surface of the electrodes, and minimizing the increase in series resistance of the trench fill with reduced-size lithographic features. Process technology and structural innovations have been introduced to overcome these limits and are discussed in the following paragraphs.

Extending the deep-trench capacitor

Process challenges

As discussed earlier, the deep-trench storage capacitor has topography advantages over the stacked capacitor; the trench is fully planarized to the silicon surface [17] and does not degrade the subsequent lithographic steps by introducing topography or high-aspect-ratio vias. Additionally, the trench is formed before the array-transistor and support-circuit CMOS, and may be subjected to high-temperature processing steps (including reduction and oxidation) without degrading the support-device performance. Furthermore, the electrical connection between the trench and the array transistor may be made without the introduction of a dedicated lithographic masking level [8], and it does not require the formation of a contact level that is borderless to the bitline wiring level. DRAM using a trench capacitor is particularly well suited for embedded-memory applications because the thermal budget associated with the capacitor occurs before the CMOS support devices are fabricated, and because the interconnect can be optimized for low-*RC* delay performance such as a copper damascene metallurgy.

The IBM trench-storage capacitor consists of a very-high-aspect-ratio contact-style hole pattern etched into the substrate, a thin storage-node dielectric insulator, a doped low-pressure chemical vapor deposition (LPCVD) polysilicon fill, and a buried-plate diffusion in the substrate. A trench capacitor at an intermediate point in the fabrication process of the cell is shown in **Figure 29**. The doped LPCVD silicon fill and the buried plate serve as the electrodes of the capacitor. A dielectric isolation collar in the upper region of the trench prevents leakage of the signal charge from the storage-node diffusion (not shown) to the buried-plate diffusion of the capacitor.

Dry-etch patterning of the deep trench has required a progression of process technology innovations, as shown in **Figure 30**. The formation of the deep trench directly into the silicon substrate was enabled by the commercial availability of reactive-ion-etch systems. The silicon is etched with high selectivity to an oxide hard mask using common halogen feed gases. Magnetically enhanced RIE at the 0.5-μm generation and dipole ring magnet RIE at 0.25-μm technologies were introduced to maintain etch-rate throughput and profile control. A hard-mask doped-oxide material that is selectively removed early in the trench process was introduced for the 0.175-μm generation. RF techniques such as dual-frequency RIE will be used to extend the silicon etching in the 0.120-μm regime. Beyond 0.1 μm, the aspect ratio (depth/width) of the trench may be limited, and the introduction of area-enhancement techniques and/or high-*k* thin dielectric insulators may be necessary.

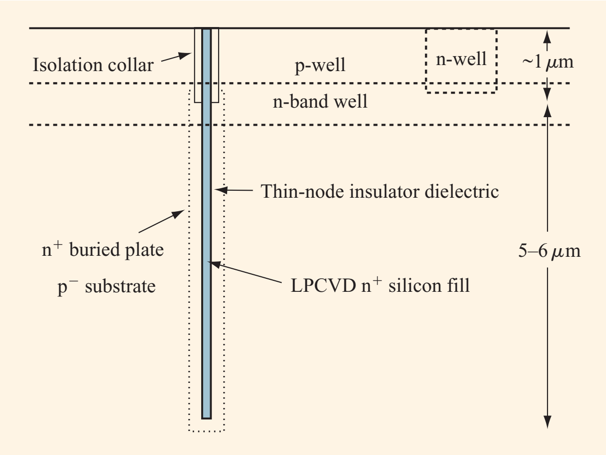

Figure 29

Cross section of the deep-trench capacitor at an intermediate point in the fabrication process of the cell.

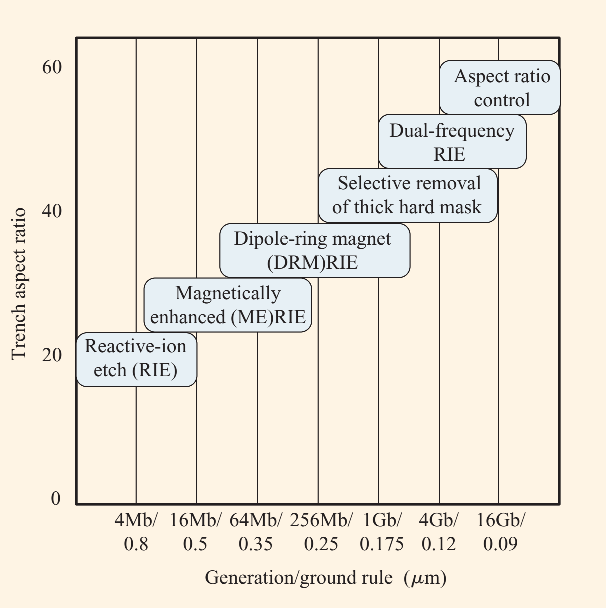

Figure 30

Progression of deep-storage-trench process technology innovations enabling etching of higher-aspect-ratio (depth/width) trenches.

Capacitance challenges

The thin dielectric insulator used for the deep-trench capacitor comprises a thermal nitridation at the substrate/plate interface, and an LPCVD silicon nitride that is subjected to a thermal oxidation to reduce leakage current and improve the dielectric reliability [50]. An H_2 anneal is used prior to the thermal nitridation to control the oxygen content at the substrate interface. Reliability projections for reoxidized nitrides support intrinsic breakdown beyond a ten-year lifetime for DRAM storage-node dielectrics as thin as 2.9 nm (equivalent oxide thickness). However, a dielectric leakage limit of 0.1 fA/μm^2 at 1 V and 85°C is reached at the 0.15-μm technology generation [50]; reductions in the dielectric thickness below ~3.0 nm result in unacceptably large dielectric leakage currents. New materials such as high-*k* dielectrics may be required to continue scaling of the

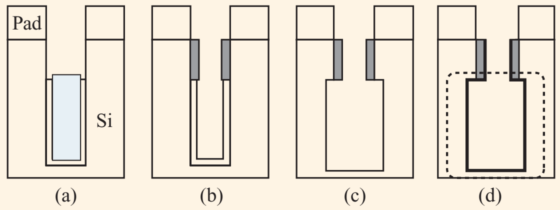

Figure 31

Area-enhancement techniques, such as the bottle-shaped capacitor, are effective in pushing trench storage technology to 0.100-μm minimum feature size: (a) Resist recess for SiN barrier definition in lower portion of trench; (b) LOCOS sidewall oxidation (isolation collar) after barrier etch and resist strip; (c) bottle enlargement using isotropic Si etching; (d) buried-plate formation self-aligned to the collar, node dielectric formation. Adapted with permission from [52]; © 1999 IEEE.

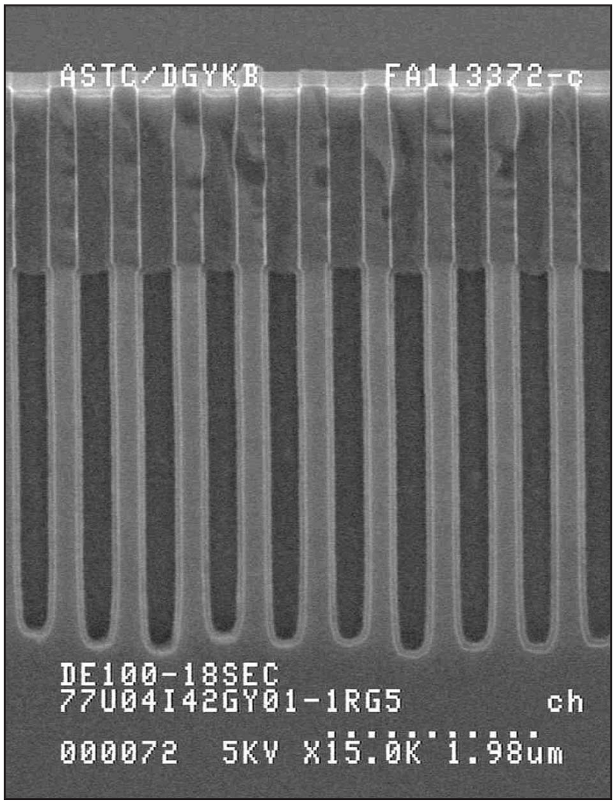

Figure 32

SEM cross section showing enlargement of the lower portion of the storage trench by use of an isotropic etch. Reprinted with permission from [52]; © 1999 IEEE.

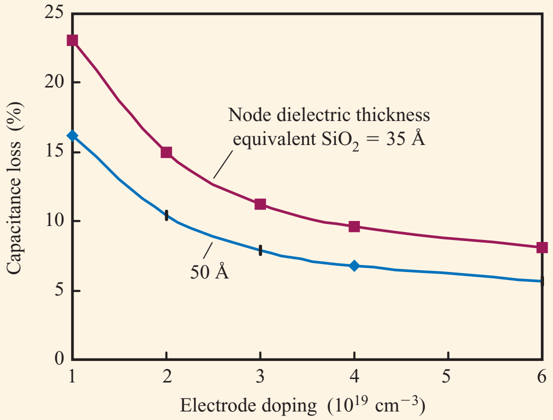

Figure 33

Percentage of capacitance loss due to depletion of majority carriers from either the surface of the buried-plate diffusion (for a stored "0") or the polysilicon in the trench (for a stored "1"). As the capacitor dielectric is thinned, the doping concentration must be increased to avoid excessive loss of capacitance. A storage node swing from 0.0 to 1.5 V with a plate voltage of 0.75 V is assumed.

effective thickness of the node dielectric into the sub-0.1-μm regime.

Although the scalability of nitride/oxide capacitor dielectrics is running into limitations, trench-capacitor cells have still been successful in advancing the technology by employing area-enhancement techniques. One example of area enhancement is the "bottle-shaped trench" [51, 52], in which the area of the trench is enhanced with an isotropic etch below the surface, as shown in **Figure 31**. An SEM cross section of trenches fabricated with this process is shown in **Figure 32**. Extrapolation of current process capability suggests that conventional dielectrics (Si_3N_4/SiO_2 or NO) will be sufficient to allow trench capacitors to be scaled to a feature size of 0.1 μm.

Capacitance loss due to majority-carrier depletion

To balance the electric field in the capacitor dielectric when a "1" or a "0" is stored, the buried-plate diffusion is customarily maintained at a constant voltage which is midway between the high and low levels stored on the capacitor; some of the reasons for seeking a balanced electric field are to maximize the reliability of the dielectric and to minimize its leakage current. Therefore, when a "0" is stored on the capacitor, the potential of the n^+ polysilicon fill is negative with respect to the n^+ buried-plate diffusion, resulting in depletion of majority carriers from the surface of the buried-plate diffusion. As storage-node dielectrics are thinned, higher buried-plate doping is needed to reduce relative capacitance loss due to the majority-carrier depletion layer. A similar capacitance loss occurs due to the depletion region in the polysilicon fill of the trench capacitor when a high level is stored. **Figure 33** illustrates the sensitivity of the capacitance loss to electrode (either buried-plate diffusion or polysilicon fill) doping concentration. The introduction of buried-plate dopant into the sidewall of the deep trench has required additional innovation in process technology. Deposition and removal of a solid doping source [53] (such as arsenic-doped glass) for the buried-plate diffusion on the lower portion of the deep trench are increasingly difficult as ground rules shrink and as trench aspect ratio increases. Gas-phase doping and plasma doping [54, 55] have been proposed to solve these problems. Regarding the effect of doping concentration in the n^+ polysilicon fill, besides capacitance loss due to depletion effects, doping limitations in the trench polysilicon pose series resistance concerns as the minimum feature size is scaled down. Therefore, as dimensions are scaled, higher doping concentration is sought in both the buried-plate diffusion and the polysilicon trench fill.

Series-resistance considerations

The series resistance of the polysilicon fill of the trench capacitor is another factor that can potentially limit continued scaling. The cross-sectional area of the trench diminishes approximately as the square of the minimum feature size. On the other hand, the trench depth remains approximately constant, or increases, from generation to generation. This causes the resistance to increase sharply. The storage-trench capacitor can be considered to be a transmission line consisting of resistance and capacitance distributed along its depth. **Figure 34** shows a simplified equivalent circuit from the bitline contact through the various contributors to series resistance and into the storage capacitor. The channel of the MOSFET, the buried strap, and the trench top region (narrowed polysilicon fill adjacent to the isolation collar, as shown in Figure 16) are represented by lumped resistances. As illustrated in **Figure 35**, during the writing of a high level (i.e., “1”) to the storage capacitor, the charge distribution on the capacitor is a function of depth in the trench. Once the access transistor is shut off, the charge on the capacitor equilibrates, and the stored voltage is lower than the high level of the bitline. In the representative case shown for a trench capacitor having a circular opening of 0.09 μm diameter and a series resistance of 86 kΩ, the voltage stored on the capacitor in a 10-ns write window is about 350 mV less than the 1.5-V bitline-high level. This

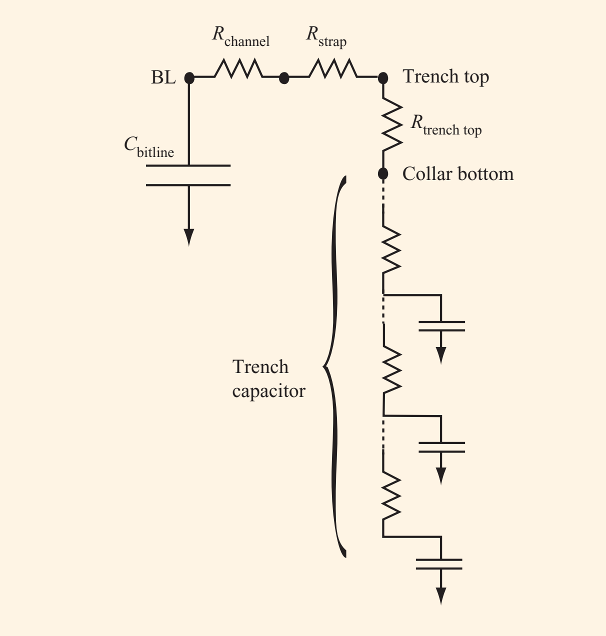

Figure 34

Simplified equivalent circuit for the trench storage capacitor and resistance components of cell. Three sections of the *RC* (resistance–capacitance) transmission-line representation of the trench storage capacitor are shown.

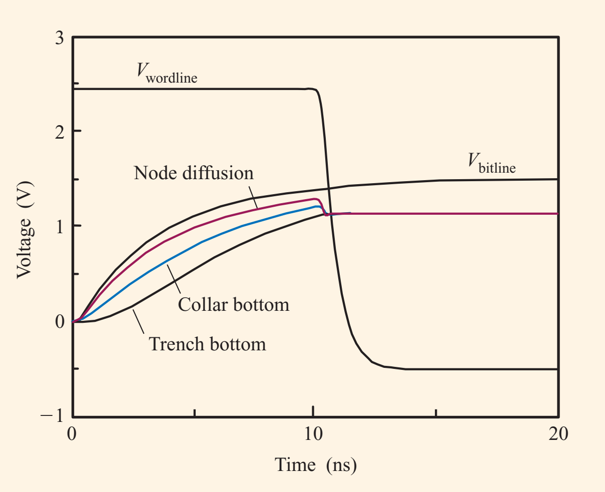

Figure 35

Simulation of representative waveforms of a write “1” operation for a polysilicon-filled storage trench with a circular cross section having a diameter of 0.090 μm. Storage-trench, strap, and trench top resistances are 86 kΩ, 20 kΩ, and 20 kΩ, respectively. Storage-trench and bitline capacitances are 33 fF and 200 fF, respectively. Of importance is the voltage on the capacitor, which is a function of depth in the trench during the charging process. The charge in the trench equilibrates such that the stored voltage is about 350 mV lower than the bitline-high voltage when the array MOSFET is shut off.

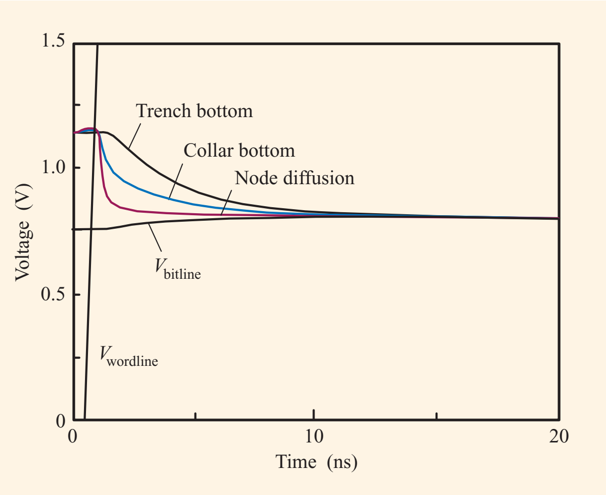

Figure 36

Simulation of representative waveforms of a read "1" operation, corresponding to the data written in Figure 35. For this polysilicon-filled trench-storage capacitor having a diameter of 0.90 μm, a 56-mV change is induced on the bitline from the 0.75-V bitline pre-charge level after 10 ns.

means that there is less charge available to be transferred between the capacitor and the bitline for a read operation. As shown in **Figure 36**, during the read operation the voltage on the capacitor is also a function of the depth in the trench. For read times less than 10 ns, the lower portion of the capacitor contributes less to the transferred charge than the upper portion, reducing the effective storage capacitance. In the case shown, only 56 mV of signal, V_{signal}, is induced on the bitline after 10 ns. For this n^+ polysilicon trench fill, at ground rules smaller than about 0.100 μm, the effect of storage-trench series resistance is quite pronounced. Novel reduced-resistivity materials will be required to extend trench capacitors to smaller ground rules.

Conclusions

Major challenges facing DRAM scaling to the 0.1-μm generation and beyond have been reviewed. Continued DRAM cost-per-bit productivity depends on reducing the cell size more rapidly than scaling minimum feature size alone would allow. Thus, cell layouts which are more compact than $8F^2$ are required, along with successful scaling of the chip area occupied by both the cell-access transistor and the storage capacitor. A crossroads on the path to continued DRAM scaling has been reached, since multiple solutions exist for these challenges. Although the best choice for continued cost-per-bit productivity is not yet clear, the most promising options have been identified and discussed.

Possible options for reducing the area occupied by the access transistor include aggressive voltage scaling for planar MOSFETs or a paradigm shift to vertical-channel MOSFETs. Each of these approaches presents significant challenges, such as very severe channel doping profile requirements for voltage-scaled planar MOSFETs, and overcoming cell-to-cell interactions that occur for certain vertical cell layouts. It appears that adoption of vertical MOSFETs in DRAM cells may provide longer-term scalability. However, the economics of manufacturability will ultimately determine the actual limits of scalability.

Regarding choice of storage capacitor, trench-storage DRAM cells promise much easier integration of vertical MOSFETs than do stacked-capacitor cells. Furthermore, a clear scaling path exists down to 0.10 μm for trench capacitors, through the use of capacitance-enhancing techniques such as the bottle-shaped trench. Scaling beyond 0.10 μm will require insulators with higher dielectric capacitance and/or trench-fill materials with lower resistance. On the other hand, the scaling path for stacked-capacitor cells is less certain because of a number of difficult challenges which appear earlier than for the trench capacitor. To obtain sufficient capacitance per cell for the stacked capacitor, geometric constraints imposed by mechanical stability and topographic considerations will force new materials to be introduced into manufacturing—first Ta_2O_5, by the 0.12-μm generation, and then BSTO. Also, the need to contain the thermal budget seen by high-performance CMOS logic integrated with DRAM favors the use of trench-capacitor storage.

Finally, as cells are made more compact than $8F^2$, the limit of the folded-bitline architecture, vertically twisted two-layer bitline wiring may have to be used to obtain the required noise immunity. The introduction of $6F^2$ and smaller cells favors trench-capacitor technology because of problems with integrating an additional level of bitline metal over stacked-capacitor cell topography.

Acknowledgments

The authors thank the members of the IBM–Infineon DRAM development team, as well as colleagues in the IBM Research Division, for valuable contributions to the development of the technology presented in this paper.

References

1. R. H. Dennard, "Field-Effect Transistor Memory," U.S. Patent 3,387,286, 1968.
2. E. Adler, J. K. DeBrosse, S. F. Geissler, S. J. Holmes, M. D. Jaffe, J. B. Johnson, C. W. Koburger III, J. B. Lasky, B. Lloyd, G. L. Miles, J. S. Nakos, W. P. Noble, Jr., S. H. Voldman, M. Armacost, and R. Ferguson, "The Evolution of IBM CMOS DRAM Technology," *IBM J. Res. & Dev.* **39,** 167–188 (1995).
3. K. Kim, C.-G. Hwang, and J. Lee, "DRAM Technology Perspective for Gigabit Era," *IEEE Trans. Electron Devices* **45,** 598–608 (1998).

4. R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, and A. R. LeBlanc, “Design of Ion-Implanted MOSFET’s with Very Small Physical Dimensions,” *IEEE J. Solid-State Circuits* **SC-9,** 256–268 (1974).
5. W. Noble and W. Walker, “Fundamental Limitations on DRAM Storage Capacitors,” *IEEE Circuits & Devices Magazine* **1,** 45–51 (1985).
6. B. Davari, C. W. Koburger, R. Schulz, J. D. Warnock, T. Furukawa, M. Jost, Y. Taur, W. G. Schwittek, J. K. DeBrosse, M. L. Kerbaugh, and J. L. Mauer, “A New Planarization Technique, Using a Combination of RIE and Chemical Mechanical Polish (CMP),” *IEDM Tech. Digest*, p. 861 (1989).
7. D. Kenney, P. Parries, P. Pan, W. Tonti, W. Cote, S. Dash, P. Lorenz, W. Arden, R. Mohler, S. Roehl, A. Bryant, W. Haensch, B. Hoffman, M. Levy, A. J. Yu, and C. Zeller, “A Buried-Plate Trench Cell for 64-Mb DRAM,” *IEEE Symposium on VLSI Technology, Digest of Technical Papers*, 1992, pp. 14, 15.
8. L. Nesbit, J. Alsmeier, B. Chen, J. DeBrosse, P. Fahey, M. Gall, J. Gambino, S. Gernhardt, H. Ishiuchi, R. Kleinhenz, J. Mandelman, T. Mii, M. Morikado, A. Nitayama, S. Parke, H. Wong, and G. Bronner, “A 0.6μm^2 256Mb DRAM Cell with Self-Aligned BuriEd Strap (BEST),” *IEDM Tech. Digest*, pp. 627–630 (1993).
9. R. H. Dennard, “Scaling Challenges for DRAM and Microprocessors in the 21st Century,” *Electrochemical Society Proceedings,* Vol. 97-3, 1997, pp. 519–532.
10. D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, “Device Scaling Limits of Si MOSFETs and Their Application Dependencies,” *Proc. IEEE* **89,** 259–288 (2001).
11. A. Hiraiwa, M. Ogasawara, N. Natsuaki, Y. Itoh, and H. Iwai, “Local-Field-Enhancement Model of DRAM Retention Failure,” *IEDM Tech. Digest*, pp. 157–160 (1998).
12. T. Hamamoto, S. Sugiura, and S. Sawada, “On the Retention Time Distribution of Dynamic Random Access Memory (DRAM),” *IEEE Trans. Electron Devices* **45,** 1300 (1998).
13. K. Yamaguchi, “Theoretical Study of Deep-Trap-Assisted Anomalous Currents in Worst-Bit Cells of Dynamic Random-Access Memories (DRAM’s),” *IEEE Trans. Electron Devices* **47,** 774 (2000).
14. H. Sunami, T. Kure, N. Hashimoto, K. Itoh, T. Toyabe, and S. Asai, “A Corrugated Capacitor Cell (CCC) for Megabit Dynamic MOS Memories,” *IEDM Tech. Digest*, pp. 806–808 (1982).
15. S. Wolf, *Silicon Processing for the VLSI Era—Volume 2: Process Integration*, Lattice Press, Sunset Beach, CA, 1990, pp. 600–615.
16. H. Kang, K. Kim, Y. Shin, I. Park, K. Ko, C. Kim, K. Oh, S. Kim, C. Hong, K. Kwon, J. Yoo, Y. Kim, C. Lee, W. Paick, D. Suh, C. Park, S. Lee, S. Ahn, C. Hwang, and M. Lee, “Highly Manufacturable Process Technology for Reliable 256 Mbit and 1 Gbit DRAMs,” *IEDM Tech. Digest*, pp. 635–638 (1994).
17. G. Bronner, H. Aochi, M. Gall, J. Gambino, S. Gernhardt, E. Hammerl, H. Ho, J. Iba, H. Ishiuchi, M. Jaso, R. Kleinhenz, T. Mii, M. Narita, L. Nesbit, W. Neumueller, A. Nitayama, T. Ohiwa, S. Parke, J. Ryan, T. Sato, H. Takato, and S. Yoshikawa, “A Fully Planarized 0.25μm CMOS Technology for 256Mbit DRAM and Beyond,” *IEEE Symposium on VLSI Technology, Digest of Technical Papers*, 1995, pp. 15, 16.
18. S. Crowder, S. Stiffler, P. Parries, G. Bronner, L. Nesbit, W. Wille, M. Powell, A. Ray, B. Chen, and B. Davari, “Trade-Offs in the Integration of High-Performance Devices with Trench Capacitor DRAM,” *IEDM Tech. Digest*, pp. 45–48 (1997).
19. S. Crowder, R. Hannon, H. Ho, D. Sinitsky, S. Wu, K. Winstel, B. Khan, S. R. Stiffler, and S. S. Iyer, “Integration of Trench DRAM into a High-Performance 0.18μm Logic Technology with Copper BEOL,” *IEDM Tech. Digest*, pp. 1017–1020 (1998).
20. H. Takato, H. Koike, T. Yoshida, and H. Ishiuchi, “Embedded DRAM Technology: Past, Present and Future,” *Proceedings of the International Symposium on VLSI Technology, Systems, and Applications*, 1999, pp. 239–242.
21. S. S. Iyer and H. L. Kalter, “Embedded DRAM Technology: Opportunities and Challenges,” *IEEE Spectrum* **36,** 56–64 (1999).
22. O. Takahashi, S. Dong, M. Ohkubo, S. Onishi, R. Dennard, R. Hannon, S. Crowder, S. Iyer, M. Wordeman, B. Davari, W. Weinberger, and N. Aoki, “1-MHz Fully Pipelined 3.7ns Address Access Time 8k×1024 Embedded Synchronous DRAM Macro,” *IEEE J. Solid-State Circuits* **35,** 1673–1679 (2000).
23. R.-P. Vollertsen and W. W. Abadeer, “Comprehensive Gate-Oxide Reliability Evaluation for DRAM Processes,” *Microelectron. Reliabil.* **36,** 1631–1638 (1996).
24. Semiconductor Industry Association (SIA), *National Technology Roadmap for Semiconductors*, 2000; see *http://www.sematech.org/public/publications/index.htm.*
25. K. Itoh, Y. Nakagome, S. Kimura, and T. Watanabe, “Limitations and Challenges of Multigigabit DRAM Chip Design,” *IEEE J. Solid-State Circuits* **32,** 624–634 (1997).
26. T. Yamagata, S. Tomishima, M. Tsukude, Y. Hashizume, and K. Arimoto, “Circuit Design Techniques for Low-Voltage Operating and/or Giga-Scale DRAMs,” *International Solid State Circuits Conference Digest of Technical Papers*, 1995, pp. 248–249, 374.
27. T. Y. Chan, J. Chen, P. K. Ko, and C. Hu, “The Impact of Gate-Induced Drain Leakage Current on MOSFET Scaling,” *IEDM Tech. Digest*, p. 719 (1987).
28. Y. Li, J. Mandelman, P. Parries, Y. Matsubara, Q. Ye, R. Rengarajan, J. Alsmeier, B. Flietner, D. Wheeler, H. Akatsu, R. Divakaruni, R. Mohler, K. Sunouchi, G. Bronner, and T. C. Chen, “Array Pass Transistor Design in Trench Cell for Gbit DRAM and Beyond,” *Proceedings of the International Symposium on VLSI Technology, Systems, and Applications*, 1999, pp. 251–254.
29. TMA TSUPREM-4, Version 6.5, 1997, Technology Modeling Associates, Inc. (acquired in January 1998 by Avant! Corporation).
30. E. Buturla, J. Johnson, S. Furkay, and P. Cottrell, “A New Three-Dimensional Device Simulation Formulation,” *NASCODE VI: Proceedings of the Sixth International Conference on Numerical Analysis of Semiconductor Devices and Integrated Circuits*, J. J. H. Miller, Ed., Boole Press Ltd., Dublin, 1989, p. 291.
31. W. Noble, S. Voldman, and A. Bryant, “The Effects of Gate Field on the Leakage Characteristics of Heavily Doped Junctions,” *IEEE Trans. Electron Devices* **36,** 720–726 (1989).
32. G. Bronner, T. Furukawa, M. Hakey, S. Holmes, D. Horak, J. Mandelman, and P. Rabidoux, “Method for Making a DRAM Cell with Grooved Transfer Device,” U.S. Patent 6,037,194, 2000.
33. D. Horak, T. Furukawa, S. Holmes, M. Hakey, W. Ma, and J. Mandelman, “Methods of Making a Trench Storage DRAM Cell Including a Step Transfer Device,” U.S. Patent 6,063,658, 2000.
34. T. Furukawa, D. Horak, S. Holmes, M. Hakey, and J. Mandelman, “DRAM Cell with Three-Sided-Gate Transfer Device,” U.S. Patent 6,121,651, 2000.
35. H.-S. Wong, K. Chan, and Y. Taur, “Self-Aligned (Top and Bottom) Double-Gate MOSFET with a 25nm Thick Silicon Channel,” *IEDM Tech. Digest*, pp. 427–430 (1997).

36. J. A. Mandelman, J. E. Barth, J. K. DeBrosse, R. H. Dennard, H. L. Kalter, J. Gautier, and H. I. Hanafi, “Floating-Body Concerns for SOI Dynamic Random Access Memory (DRAM),” *Proceedings of the IEEE International SOI Conference*, September 1996, pp. 136–137.
37. W. F. Richardson, D. M. Bordelon, G. P. Pollack, A. H. Shah, S. D. S. Malhi, H. Shichijo, S. K. Banerjee, M. Elahy, R. H. Womack, C.-P. Wang, J. Gallia, H. E. Davis, and P. K. Chatterjee, “A Trench Transistor Cross-Point DRAM Cell,” *IEDM Tech. Digest*, pp. 714–717 (1985).
38. U. Gruening, C. J. Radens, J. A. Mandelman, A. Michaelis, M. Seitz, N. Arnold, D. Lea, D. Casarotto, A. Knorr, S. Halle, T. H. Ivers, L. Economikos, S. Kudelka, S. Rahn, H. Tews, H. Lee, R. Divakaruni, J. J. Welser, T. Furukawa, T. S. Kanarsky, J. Alsmeier, and G. B. Bronner, “A Novel Trench DRAM Cell with a Vertical Access Transistor and Buried Strap (VERI BEST) for 4Gb/16Gb,” *IEDM Tech. Digest*, pp. 25–29 (1999).
39. K. Itoh, “Trends in Megabit DRAM Circuit Design,” *IEEE J. Solid-State Circuits* **25,** 778–789 (1990).
40. Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, Cambridge, UK, 1998.
41. C. Radens, U. Gruening, J. Mandelman, M. Seitz, T. Dyer, D. Lea, D. Casarotto, L. Clevenger, L. Nesbit, R. Malik, S. Halle, S. Kudelka, H. Tews, R. Divakaruni, J. Sim, A. Strong, D. Tibbel, N. Arnold, S. Bukofsky, J. Preuninger, G. Kunkel, and G. Bronner, “A 0.135μm^2 6F^2 Trench-Sidewall Vertical Device Cell for 4Gb/16Gb DRAM,” *IEEE 2000 Symposium on VLSI Technology, Digest of Technical Papers*, 2000, pp. 80–81.
42. D. Takashima, S. Watanabe, H. Nakano, Y. Oowaki, and K. Ohuchi, “Open/Folded Bit-Line Arrangement for Ultra-High-Density DRAM’s,” *IEEE J. Solid-State Circuits* **29,** 539–542 (1994).
43. T. Hamada, N. Tanabe, H. Watanabe, K. Takeuchi, N. Kasai, H. Hada, K. Shibahara, K. Tokashiki, K. Nakajima, S. Hirasawa, E. Ikawa, T. Saeki, E. Kakehashi, S. Ohya, and T. Kunio, “A Split-Level Diagonal Bit-Line (SLDB) Stacked Capacitor Cell for 256Mb DRAMs,” *IEDM Tech. Digest*, pp. 799–802 (1992).
44. H. Hidaka, Y. Matsuda, and K. Fujishima, “A Divided/Shared Bit-Line Sensing Scheme for ULSI DRAM Cores,” *IEEE J. Solid-State Circuits* **26,** 473–478 (1991).
45. T. Kirihata, G. Mueller, M. Clinton, S. Loeffler, B. Ji, H. Terletzki, D. Hanson, C. Hwang, G. Lehmann, D. Storaska, G. Daniel, L. Hsu, O. Weinfurtner, T. Boehler, J. Schnell, G. Frankowsky, D. Netis, J. Ross, A. Reith, O. Kiehl, and M. Wordeman, “A 113μm^2 600Mb/s/pin 512Mb DDR2 SDRAM with Vertically-Folded Bitline Architecture,” *International Solid State Circuits Conference, Digest of Technical Papers*, 2001, pp. 382–383, 468.
46. H. Hoenigschmid, A. Frey, J. DeBrosse, T. Kirihata, G. Mueller, D. Storaska, G. Daniel, G. Frankowsky, K. Guay, D. Hanson, L. Hsu, B. Ji, D. Netis, S. Pararoni, C. Radens, A. Reith, H. Terletzki, O. Weinfurtner, J. Alsmeier, W. Weber, and M. Wordeman, “A 7F^2 Cell and Bitline Architecture Featuring Tilted Array Devices and Penalty-Free Vertical BL Twists for 4-Gb DRAM’s,” *IEEE J. Solid-State Circuits* **35,** 713–718 (2000).
47. C. Radens, U. Gruening, M. Weybright, J. DeBrosse, R. Kleinhenz, H. Hoenigschmid, A. Thomas, J. Mandelman, J. Alsmeier, and G. Bronner, “A 0.21μm^2 7F^2 Trench Cell with a Locally-Open Globally-Folded Dual Bitline for 1Gb/4Gb DRAM,” *IEEE Symposium on VLSI Technology, Digest of Technical Papers*, 1998, pp. 36–37.
48. H. Nakano, D. Takashima, K. Tsuchida, S. Shiratake, T. Inaba, M. Ohta, Y. Oowaki, S. Watanabe, K. Ohuchi, and J. Matsunaga, “A Dual Layer Bitline DRAM Array with Vcc/Vss Hybrid Precharge for Multi-Gigabit DRAMs,” *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, 1996, pp. 190–191.
49. H. Hidaka, K. Fujishima, Y. Matsuda, M. Asakura, and T. Yoshihara, “Twisted Bit-Line Architectures for Multi-Megabit DRAM’s,” *IEEE J. Solid-State Circuits* **24,** 21–27 (1989).
50. E. Wu, C. Hwang, R. Vollertsen, H. Shen, R. Kleinhenz, C. Radens, and A. Strong, “Thickness and Polarity Dependence of Intrinsic Breakdown of Ultra-Thin Reoxidized Nitride for DRAM Technology Applications,” *IEDM Tech. Digest*, pp. 77–80 (1997).
51. M. Schrems, J. Mandelman, J. Hoepfner, H. Schaefer, and R. Stengl, “Bottle-Shaped Trench Capacitor with Epi Buried Layer,” U.S. Patent 6,018,174, 2000.
52. T. Rupp, N. Chaudary, K. Dev, Y. Fukuzaki, J. Gambino, H. Ho, J. Iba, E. Ito, E. Kiewra, B. Kim, M. Maldei, T. Matsunaga, J. Ning, R. Rengarajan, A. Sudo, Y. Takegawa, D. Tobben, M. Weybright, G. Worth, R. Divakaruni, R. Srinivasan, J. Alsmeier, and G. Bronner, “Extending Trench DRAM Technology to 0.15μm Groundrule and Beyond,” *IEDM Tech. Digest*, pp. 33–36 (1999).
53. L. Economikos, C. Murthy, and R. Young, “Study of Arsenic Out-Diffusion for Buried Plate Formation in Trench Capacitors,” *Proceedings of the IEEE/CPMT International Electronics Manufacturing Technology Symposium*, 1998, pp. 423–432.
54. S. Saida, T. Sato, I. Mizushima, Y. Ozawa, and Y. Tsunashima, “Single Layer Nitride Capacitor Dielectric Film and High Concentration Doping Technology for 1Gb/4Gb Trench-Type DRAMs,” *IEDM Tech. Digest*, pp. 265–268 (1997).
55. K. Lee, B. Lee, J. Hoepfner, L. Economikos, C. Parks, C. Radens, J. Bernstein, and P. Kellerman, “Plasma Immersion Ion Implantation as an Alternative Deep Trench Buried-Plate Doping Technology,” *Proceedings of the Thirteenth International Conference on Ion Implantation Technology*, 2000, pp. 460–463.

Received May 30, 2001; accepted for publication October 9, 2001

Jack A. Mandelman *IBM Microelectronics Division, East Fishkill facility, Route 52, Hopewell Junction, New York 12533 (modelman@ieee.org).* Dr. Mandelman received the B.E.E. and M.E.E. degrees from the City College of New York in 1969 and 1971, respectively, and the Ph.D.E.E. degree from the City University of New York in 1975. He subsequently joined IBM in Burlington, Vermont, as a circuit designer in a 32Kb DRAM program. Since then, Dr. Mandelman's primary focus has been on the application of simulation to device design and process integration of advanced DRAM and logic technologies. His contributions, which span more than eight generations of IBM MOS technology, involve the areas of device reliability, SOI floating-body effects, parasitic leakage mechanisms in trench isolation, and novel memory cell designs. In 1992 Dr. Mandelman joined the IBM/Siemens/Toshiba 256Mb DRAM Development Alliance at the IBM Semiconductor Research and Development Center in Hopewell Junction, New York. Most recently, he has driven the paradigm shift from planar to vertical MOSFET DRAM. He is currently a Senior Engineer and IBM's lead device designer for advanced vertical-MOSFET DRAM technology for 1 Gb and beyond. Dr. Mandelman is one of IBM's most decorated inventors, holding more than 150 U.S. patents with more than 100 pending. He has received numerous IBM valuable-patent awards for innovations in DRAM cell structure, process integration, and SOI technology. At present he is a senior business counsel for Intellectual Property and chairs the IBM device and circuit invention review board. Dr. Mandelman is a Senior Member of the IEEE.

Robert H. Dennard *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (dennard@us.ibm.com).* Dr. Dennard is an IBM Fellow at the IBM Thomas J. Watson Research Center, Yorktown Heights, New York. He received B.S. and M.S. degrees in electrical engineering from Southern Methodist University in 1954 and 1956 and the Ph.D. degree from Carnegie Institute of Technology in 1958. He subsequently joined the IBM Research Division, where he has been involved in microelectronics research and development from its early days. In 1967 he invented the one-transistor dynamic memory cell (DRAM) used in most computers today. With co-workers, he developed the concept of MOSFET scaling in 1972. Dr. Dennard is a Fellow of the IEEE and received their Edison Medal in June, 2001. He is a member of the National Academy of Engineering and the American Philosophical Society. His honors include the National Medal of Technology, presented by President Reagan in 1988, and induction into the National Inventors' Hall of Fame in 1997.

Gary B. Bronner *IBM Microelectronics Division, East Fishkill facility, Route 52, Hopewell Junction, New York 12533 (gbronner@us.ibm.com).* Dr. Bronner joined IBM in 1985 and is currently the DRAM alliance program manager, 1Gb DRAM project manager, and an IBM Distinguished Engineer. He received a B.S. degree in electrical engineering from Brown University and M.S. and Ph.D. degrees from Stanford University. Subsequently he joined IBM at the Thomas J. Watson Research Center and initially worked on CMOS and DRAM technology. In 1989 he became Manager of DRAM Technology in IBM Research and led a joint Yorktown/Burlington team that defined the 0.25-micron technology that became the basis of the IBM/Siemens/Toshiba DRAM Development Alliance in 1993. In 1993 he transferred to the IBM Microelectronics Division to continue his role managing the DRAM development team. In this capacity, he has helped define and deliver the last five generations of DRAM technology inside IBM. He has participated in several IBM Academy of Technology study groups and organized and co-chaired an IBM Academy workshop on merged DRAM/logic. He is a Senior Member of the Institute of Electrical and Electronics Engineers and was the technical program chair of the IEDM Conference in 1998 and the general chair of that conference in 1999. Dr. Bronner organized the first International Symposium on ULSI Process Integration at the Fall 1999 Electrochemical Society Meeting. He has reached the 20th IBM Invention Achievement Plateau, with 27 patents issued and 31 patents pending. He has published 20 technical disclosures and 45 articles in refereed journals. Dr. Bronner has received a Research Division Outstanding Contribution Award, a Division Award, an Outstanding Technical Achievement Award, a Division Excellence Award, and a Division Patent Portfolio Award.

John K. DeBrosse *IBM Microelectronics Division, Burlington facility, 100 River Street, Essex Junction, Vermont 05452 (jdebros@us.ibm.com).* Mr. DeBrosse received a B.S.E.E. degree in 1983 and an M.S.E.E. degree in 1984, both from Purdue University. In 1985 he joined IBM, where he has contributed to five generations of DRAM technology development and product design (4Mb–1Gb). He is an author or co-author of 20 patents and 18 technical papers, and is currently involved in MRAM product design.

Rama Divakaruni *IBM Microelectronics Division, East Fishkill facility, Route 52, Hopewell Junction, New York 12533 (rdivakar@us.ibm.com).* Dr. Divakaruni received a B.Tech. degree in electrical engineering from the Indian Institute of Technology, Madras, in 1988 and a Ph.D. degree in electrical engineering from the University of California at Los Angeles in 1994. He subsequently joined the IBM Microelectronics Division in Essex Junction, Vermont, where he worked on process characterization of 16Mb DRAM shrink technologies and led the team that successfully installed IBM 0.25-μm DRAM technology in manufacturing. In early 1998 Dr. Divakaruni moved to the IBM Semiconductor Research and Development Center at the East Fishkill facility, where he was the lead integrator for the 0.15-μm DRAM technology platform developed in the IBM/Siemens/Toshiba DRAM Development Alliance. Since 1999, he has managed the vertical-pass-transistor DRAM process integration group which has demonstrated sub-$8F^2$ and $8F^2$ cell concepts scalable below 100 nm. Dr. Divakaruni holds 17 patents, with more than 60 pending. He has been an author or co-author on several papers related to vertical-pass gate transistors for DRAMs.

Yujun Li *IBM Microelectronics Division, East Fishkill facility, Route 52, Hopewell Junction, New York 12533 (yujun@us.ibm.com).* Dr. Li received a B.S. degree in physics and electrical engineering from the University of Science and Technology of China in 1992, and M.S., M.Phil., and Ph.D. degrees in electrical engineering from Yale University in 1993, 1994, and 1997, respectively. She subsequently joined the IBM Semiconductor Research and Development Center at the East Fishkill facility, where she has worked on array device design of high-density, high-performance DRAM technologies. In 1998, she received an IEEE Paul Rappaport Award for best paper in an EDS publication during that year. Dr. Li is currently an Engineering Manager in the Device Design Department of the IBM/Siemens/Toshiba DRAM Development Alliance.

Carl J. Radens *IBM Microelectronics Division, East Fishkill facility, Route 52, Hopewell Junction, New York 12533 (radens@us.ibm.com).* Dr. Radens is a Senior Engineer at the IBM Semiconductor Research and Development Center working in the area of process integration. He received a B.A. degree in physics from Oberlin College in 1983 and a Ph.D. degree in electrical engineering from the University of Cincinnati in 1990. Since joining IBM in 1990, he has worked in areas including dry-etch process development, integration of self-aligned contacts, and BEOL interconnect, storage capacitor, and DRAM cell design. He led the introduction of the trench-sidewall vertical-transistor DRAM at IBM, and is currently lead engineer for the 90-nm ground-rule DRAM program. Dr. Radens has served on the IBM Semiconductor Equipment Council dry-etch tool committee and is a member of the Integrated Circuits and Manufacturing subcommittee for the IEEE International Electron Devices Meeting. He has reached the 32nd IBM Invention Achievement Plateau, holds more than 50 U.S. patents, and is the author or coauthor of more than 25 technical publications.