The evolution of IBM CMOS DRAM technology

by 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
R. Ferguson

The development of DRAM at IBM produced many novel processes and sophisticated analysis methods. Improvements in lithography and innovative process features reduced the cell size by a factor of 18.8 in the time between the 4Mb and 256Mb generations. The original substrate plate trench cell used in the 4Mb chip is still the basis of the 256Mb technology being developed today. This paper describes some of the more important and interesting innovations introduced in IBM CMOS DRAMs. Among them, shallow-trench isolation, I-line and deep-UV (DUV) lithography, titanium salicidation, tungsten stud contacts, retrograde n-well, and planarized back-end-of-line (BEOL) technology are core elements of current state-of-the-art logic technology described in other papers in this issue. The DRAM specific features described are borderless contacts, the trench capacitor, trench-isolated cell devices, and the "strap." Finally, the methods for study and control of leakage mechanisms which degrade DRAM retention time are described.

Introduction

The 4Mb DRAM generation saw a revolutionary change in technology at IBM, with the introduction of CMOS, trench capacitor storage, and other new processes and structures. Although rapid progress continues, the basic cell structures and many of the processes developed then are being used in the 64Mb and 256Mb DRAMs being developed today. In addition, much of the technology developed for the 4Mb and 16Mb DRAMs is now used in CMOS logic technology. This paper describes the DRAM cell used by IBM beginning with the 4Mb generation, and traces its evolution to the 256Mb cell being developed today. We then describe the development of some key technology elements, and explain how key DRAM device problems were solved.

Dynamic random access memory has been a good vehicle for technology development, because there is a predictable demand for a large number of chips of standard design. The density of the array, a well-understood benchmark which determines cost, is a very effective driver of technology development. The addressability and repetitive character of the array make it possible to find and solve technology problems in the product. The high volume allows employment of the team of experts required to do a thorough development job. Thus, DRAM is the

©Copyright 1995 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

0018-8646/95/$3.00 © 1995 IBM

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

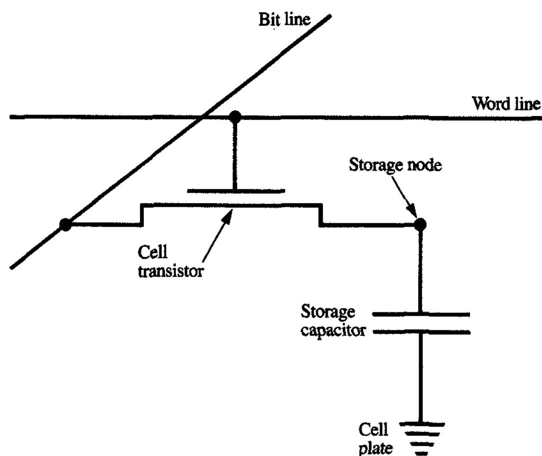

Figure 1

The one-transistor cell, consisting of a storage capacitor and a single transistor through which it is accessed.

product that has driven the state of the art of silicon device technology up to the present day.

The one-device DRAM cell [1], invented at IBM by R. Dennard, consists of a cell transistor with the drain connected to one node of the cell storage capacitor, the source connected to a bit line, and the gate connected to the word line, which runs orthogonal to the bit line (**Figure 1**). The requirement to have a large capacitor in a small space with low leakage is the main driver of DRAM technology. A brief description of the cell operation will help to explain why. To write, the bit line is driven to a high or low logic level with the cell transistor turned on, and then the cell transistor is shut off, leaving the capacitor charged high or low. Since charge leaks off the capacitor, a maximum refresh interval is specified. To read, or refresh the data in the cell, the bit line is left floating when the cell transistor is turned on, and the small change in bit-line potential is sensed and amplified to a full logic level. The ratio of cell capacitance to bit-line capacitance, called the transfer ratio, which ranges from about 0.1 to 0.2, determines the magnitude of the change in bit-line potential. A large cell capacitance is needed to deliver an adequate signal to the sense amplifier.

The evolution of technology has followed the following overall trends.

DRAM cell size has decreased from 11.3 μm^2 for the first 4Mb cell to 0.6 μm^2 for the first 256Mb cell. Improvements in lithography were responsible for much but not all of the size reduction. New process features were also necessary to shrink the cell and to improve array performance. As a result, there has been a trend toward increased process complexity, as reflected in the number of masking steps used in the process, which has increased from 13 in the 4Mb generation to 25 in the 256Mb generation.

The increase in the complexity of DRAM technology has driven up the cost of DRAM development, resulting in the formation of alliances between companies to reduce the expense to individual companies. The IBM 64Mb DRAM is being developed by an alliance between IBM and Siemens, and the 256Mb by a triple alliance including IBM, Toshiba, and Siemens.

DRAM external power supplies follow industry standards. Because the chip power is low enough, DRAM can use on-chip power supply regulation to reduce the internal circuit power supply swings. IBM has led the industry in reduction of power supply voltages for CMOS logic and memory. DRAM technology resists power supply scaling more than logic technology because of the need for storage of charge. **Table 1** illustrates this trend.

Power supply voltage reduction will come rapidly, since the market for battery-operated equipment is growing faster than previously anticipated. Also, performance competition in microprocessors demands ever-shorter channel lengths, which in turn require reduction in power because of device scaling. DRAM chips and technology will be similarly forced to operate at lower power supply voltages in the near future.

The market for battery-operated equipment also creates a need for longer DRAM retention times, to reduce the power associated with refreshing the data. The data retention time specification is currently 64–256 ms, making very low leakage current a requirement, along with a large cell capacitance.

The cell capacitor was a simple planar structure through the 1Mb generation. At and beyond the 4Mb generation, as the cell size decreased, the effective surface area of the capacitor was maintained by placing the capacitor on the sides of a narrow trench etched into the silicon, or by putting the capacitor on top of the other elements of the cell.

The next section begins by explaining the development of the IBM 4Mb substrate plate trench (SPT) DRAM cell, at a cell size of 11.3 μm. We next show how important features were added for succeeding generations to reduce the cell size to 0.6 μm^2, where it now stands for the 256Mb chip. Succeeding sections trace in more detail the development of certain technology elements essential to DRAM. We start with the strap connection between the storage trench polysilicon and the node diffusion, a unique SPT DRAM requirement, which is a challenge for process integration. Then we discuss device isolation, retrograde n-well, salicidation, lithography, and metallization. Finally, solutions for various cell device design and retention time problems encountered during DRAM development are

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

described. Included are gate-induced drain leakage (GIDL), three-dimensional device effects, dislocation-related leakage, and the variable retention time phenomenon.

DRAM cell structure evolution

The folded bit-line cell array configuration (**Figure 2**) has been used universally in the industry since the 1Mb time frame. In the folded bit-line configuration, a cell is crossed by two word lines and one bit line. One of the word lines (WL1 in Figure 2) is the "active word line" for the cell, and forms the gate of the cell device. The second word line (WL2), the "passing word line," is the gate of the cell device on the adjacent cell. Thus, the bit line (BL) and reference bit line ($\overline{BL}$) can be adjacent, leading to better matching and noise rejection, as well as providing a wider pitch for the layout of the sense amplifier. Although the cell now contains two word lines (active and passing), this does not require more cell area than an open bit-line cell, since the additional area is also generally the same area used for the storage capacitor.

IBM adopted CMOS technology for DRAM at the 4Mb generation. Previously, DRAM had been implemented in simple n-MOS technology because the latter was relatively inexpensive. However, logic applications, which were sensitive to active power, had already migrated to CMOS. Integration of a DRAM cell structure into a CMOS technology brought with it some fundamental issues which had to be resolved before tackling the cell structure in detail. For an integrated DRAM technology, the doping types and profiles from the starting substrate up to the device gates had to be chosen and optimized to the best trade-off of cost, reliability, function, and speed. The most fundamental issue, however, was the one of choice between n-well CMOS on a p-type substrate and p-well on an n-type substrate. Two important conditions were set which the technology had to meet, and which still obtain for current and future generations:

1. The array must be isolated from the substrate by building it within a well of opposite doping to take full advantage of CMOS. This benefits cell retention time by eliminating all leakage current sources associated with the substrate wafer. It reduces the incidence of soft errors due to ionizing radiation by confining the effective minority carrier collection length within the well and sending many of the generated minority carriers to the substrate, where they do not affect the storage node diffusion.
2. The well potential must be stable despite the impact ionization that accompanies FET operation. This ionization is largest for an n-channel device. Therefore, n-MOS devices should not be positioned in a well whose conductivity is reduced by light doping or constrained depth. This constraint is satisfied by an n-well CMOS technology on a heavily doped p-type substrate.

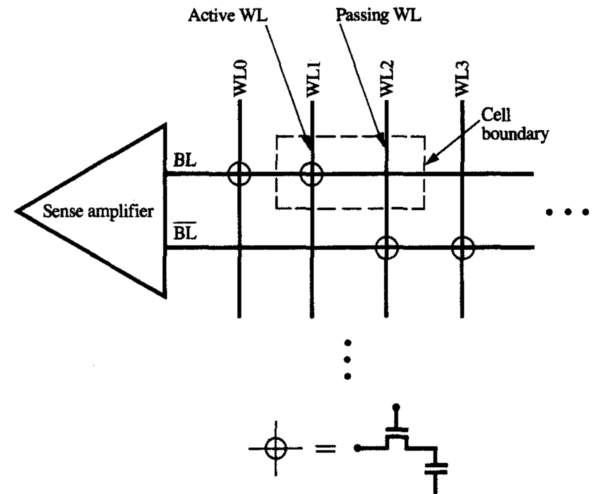

Figure 2

Folded bit-line cell configuration, which places two adjacent bit lines on the same sense amplifier.

Table 1 Power supplies by memory generation.

Memory generation	*Memory PS (V)*		*Logic PS (V)*
	External	*Internal*	
4 Mb	3.6, 5	3.6	5, 3.6
16 Mb	3.3, 5	3.3, 3.6	5, 3.6, 3.3, 2.5
64 b	3.3	3.3	3.6, 3.3, 2.5
256 Mb	3.3, 2.5	2.5	3.3, 2.5, 1.8

PS = power supply voltage

A p-MOS DRAM array built in an n-well CMOS technology meets these conditions and was chosen for the 4Mb generation. The cell choice was then made within that framework.

The criteria for cell choice are density, process simplicity, adequate storage capacitance for detectable signal, and low parasitic capacitances for performance and minimization of noise. Each generation of DRAM must compete with prior generations by providing an ultimate lower cost per bit. This is accomplished by decreasing cell size with each generation, while minimizing the increase in processing cost. The industry trend [2] is to reduce cell

DOCKET ALARM
Find authenticated court documents without watermarks at docketalarm.com.

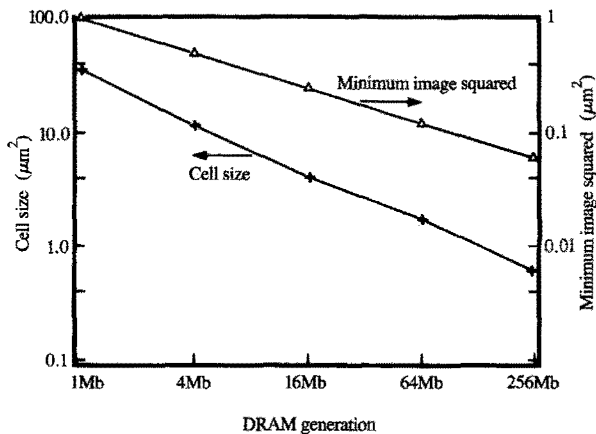

Figure 3

DRAM cell size and square of minimum lithographic image vs. generation.

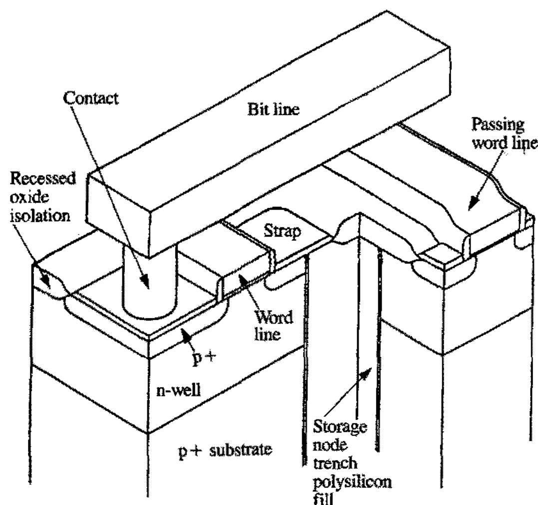

Figure 4

Cross section of 4Mb substrate plate trench (SPT) DRAM cell.

size by a factor of 0.33 for each generation. The industry trend in lithography is to reduce the minimum image size by a factor of 0.7 for each generation, so the use of lithography alone would reduce the cell area by 50% for each generation. **Figure 3** shows the cell size vs. generation plotted together with the square of the minimum lithographic image. This shows that technological innovation, involving a change in cell structure, is needed in addition to lithographic scaling to reduce the cell size by a multiple of one third for each generation. Also, technical advances are required to implement dimensional scaling (reduced heat cycles, film thicknesses, defect levels, etc.) and to mitigate electrical limitations arising from such scaling.

At the transition from 1Mb to 4Mb [3], planar capacitors did not provide enough cell capacitance, and were replaced by three-dimensional capacitors throughout the industry. These took the form of either trench capacitors buried within etched holes in the silicon [4–7] or stacked capacitors built above the silicon [8–12] in the region of the interconnect-level films.

The planar capacitor in the 1Mb and prior generations in IBM used an oxide/nitride/oxide (ONO) storage insulator consisting of a sandwich of thermally grown oxide, followed by deposited silicon nitride, which is subjected to oxidation to seal any weak spots in the nitride. Early experiments with deep-trench capacitors produced excellent results using the same ONO storage node insulator used in the 1Mb generation. Since the defect levels per unit area were much lower than predicted by experience with planar capacitors, trench capacitors were chosen for the 4Mb DRAM generation.

• *The 4Mb generation*

The cross section of the IBM 4Mb cell is shown in **Figure 4**. The capacitor consists of the polysilicon storage node electrode which fills the trench, the ONO node dielectric on the trench walls, and the p+ substrate which forms the storage plate. Thus, there is no need for the separate plate wiring layer found in other cell types. The trench polysilicon node is connected to the array device diffusion pocket by a selective silicon epitaxy surface strap, which bridges the thin oxide separating the active area and the top surface of the storage node. This cell structure is referred to as the substrate plate trench (SPT) cell [13]. This type of cell differs from the standard industry trench cells, which either form the storage node in the silicon substrate outside the trench, or stack two polysilicon electrodes separated by the insulator inside the trench.

Active device areas are formed in a p-epitaxial layer grown on the p+ substrate. As shown in the layout of **Figure 5(a)**, the active regions are separated by conventional isolation. Because the cell is in a well, a vertical parasitic p-FET is formed between the p+ storage node diffusion and the p+ substrate, with the trench polysilicon as the gate. This parasitic device is never turned on because the gate is tied to the p+ storage node diffusion, which is always the source of the p-FET, and the array n-well is back-biased at about 1 V above the power supply voltage.

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

- *The 16Mb generation*

In the 4Mb cell, a localized oxidation of silicon (LOCOS) isolation region must separate a trench from an adjacent active device area to avoid parasitic sidewall currents gated by the storage node polysilicon, and the automatic strapping of all adjacent nodes and trenches which would otherwise occur. In the 16Mb cell, this limitation was overcome by a modification of the trench structure. **Figure 6** is a cross section of the 16Mb cell, showing that the insulator lining the trench now contains a thick (approximately 100-nm) SiO_2 collar which extends from the silicon surface to a point below the n-well. The thick SiO_2 collar prevents unwanted bridging of exposed node silicon and diffusion surfaces. It also has the function of isolating the node trench polysilicon from the cell device edge under the word line, which was the role of the LOCOS isolation in the 4Mb cell. To further isolate the storage trench polysilicon from the abutting cell device region and the word line, the top of the trench polysilicon must also be recessed below the active device area wafer surface and covered by a thick oxide. The storage trench can now be placed in the space between cell devices, as shown in **Figure 5(b)**. This increases the efficiency of the cell layout by decreasing the area devoted to thick oxide isolation and increasing the area available for storage capacitance.

Electrical connection between the trench polysilicon node and the array device across the thick collar is made by a deposited polysilicon surface strap using a novel process to be described in a subsequent section of this paper. This strap is borderless to the dielectric-encapsulated word line. This reduces the active-to-passing word-line space, which was determined by the overlay tolerance of the trench, isolation, and word-line layers in the 4Mb cell. The 16Mb cell is referred to as the merged isolation and node trench (MINT) SPT cell [14].

- *The 64Mb generation*

Along with the density increases, improvements in performance were also realized as a consequence of scaling. During the 4Mb and 16Mb generations, the lower performance of a p-MOS cell device relative to n-MOS was not a problem. With the 64Mb generation, the time required to move data in and out of cells could be significant. Therefore, an n-MOS array was desired. The simplest structural change to achieve this would be simply to interchange n-material for p-material relative to the 4Mb and 16Mb generations. Thus, the starting material would be n-type, with implanted p-wells in which the cell arrays would be formed. However, this structure forfeited the noise immunity advantages of n-well technology as argued for the 4Mb and 16Mb generations. The benefits of an n-well CMOS technology on a p-type substrate could be retained at the cost of some increased process complexity. The array p-well and the substrate would have to be

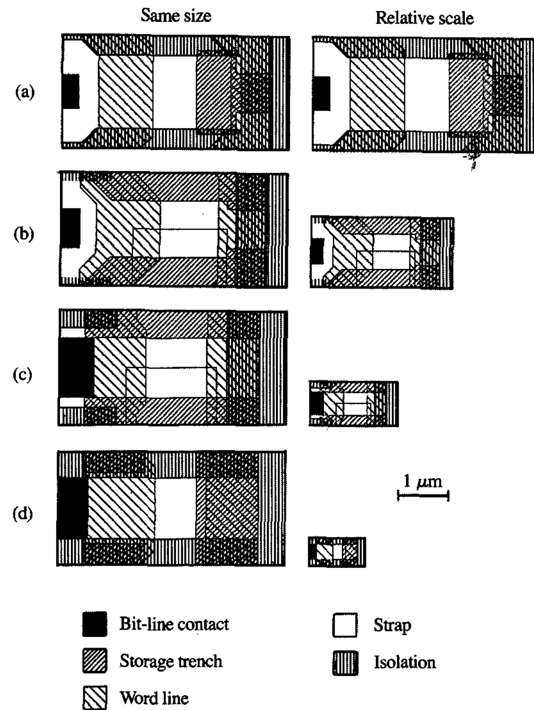

Figure 5

Layouts of DRAM cells: (a) 4Mb, (b) 16Mb, (c) 64Mb, and (d) 256Mb. Layouts are shown in both same-size and scaled-size drawings.

electrically isolated. This allowed the array well to be reverse-biased (−1 V) for low leakage, low parasitic capacitance, and maximum signal, while the substrate was at ground for low noise and best performance.

Figure 7 shows the cell configuration which achieves this for the 64Mb generation. The array p-well is isolated from the substrate by an underlying n-type layer which is formed by outdiffusion from a source deposited within the trenches. In a dense array, the trenches are close enough together that diffused regions form a continuous n-type layer. Since the n-type region extends to the bottom of the trenches, it also serves as a capacitor plate. Connection of this n-type plate to the top surface is formed by an n-well ring which surrounds the array. This cell configuration is called the buried plate trench (BPT) cell [15].

The overall cell layout is similar to that of the 16Mb generation, as shown by **Figure 5(c)**, with the addition of a "borderless contact." This feature reduces the cell size by eliminating the diffusion border required between the bit-line contact and the adjacent word line. This requires a special contact structure made by imposing a film

DOCKET ALARM Find authenticated court documents without watermarks at docketalarm.com.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS

Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS

Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS

Sync your system to PACER to automate legal marketing.

WHAT WILL YOU BUILD? | sales@docketalarm.com | 1-866-77-FASTCASE

fastcase®
Smarter legal research.