transistor DRAM cell, since higher capacitance was required to reduce the soft errors due to noise electrons generated by alpha particles from the package materials of the chip, cosmic rays and other noise sources [44]; a dual dielectric for the charge storage capacitor, utilizing the higher dielectric constant of silicon nitride formed by chemical vapor deposition (CVD) on thermally grown silicon dioxide to enhance the composite storage medium’s dielectric constant and to reduce pinholes in the thinner silicon dioxide (not all DRAM manufacturers utilized this option); plasma etch technology to produce steeper walls or trenches to reduce tapered structures which take up silicon real estate (chip area) and an optical wafer stepper to reduce the design rules from three to less than two microns. Rideout [296] and Chatterjee [299] have also reviewed these DRAM advances.

The 256K DRAM further reduced the design rule to the 1.5-2 μm range and introduced refractory metal silicides to reduce the interconnect wiring delay [44] and aluminum metal for double and triple polysilicon technologies. The M-bit DRAM era, initially a shrink of the original 2 μm 256K DRAM design, approached 1 μm design rules (see Table 2); more importantly, however, was the introduction of two three-dimensional (3-D) trench charge storage capacitors (see Figures 9 and 10). Sah has noted that the goal of these 3-D capacitor designs was to reduce the planar area of the storage capacitor while maintaining the storage capacitance at more than 32 fF to hold more than 10⁶ electrons at a VDD of 5V to limit soft errors [44]. In the stack capacitor design, multilayers of conductors (poly Si or Al) and insulators (silicon dioxide and silicon nitride) are stacked on top of the pass transistor. In the trench capacitor design, a trench is etched in the silicon and an MOS storage capacitor is fabricated in the trench, adjacent to the pass transistor which remains on the planar surface. In this case, the trench depth is about 10 μm and the spatial area is about 6-9 μm². Chatterjee and colleagues at Texas Instruments introduced a structure which placed the pass transistor inside the trench to further conserve silicon real estate [300,301].

The 4M DRAM era introduced the sub-micron design rule regime at 0.8 μm with 3-D storage capacitors. The types and features of storage cell designs have subsequently proliferated [44,302,303]. The decreasing design rules result in higher speed and reduced power-delay product as a result of lower capacitance and current [44]. The power-delay product is additionally reduced by reducing VDD [44].

The DRAM became the test vehicle par excellence to advance the silicon IC process technology because of its repetitive memory structure. In more recent years, however, especially after the U.S. makers retreated from a significant position in the manufacture of DRAMS, their expertise in the fabrication of microprocessors has propelled the logic and microprocessor family as test vehicle drivers. Nevertheless, the DRAM continues to drive the extendibility of personal computers (PCs) vis-à-vis the memory content.

**Integrated Circuit Scaling**

Gordon Moore’s remarkably prescient assessment of memory component growth in 1965, initially based on bipolar and then MOS memory density, observed that a semilog graph of the number of bits on a memory IC versus the date of initial availability was a straight line, representing almost a doubling per year [50-53]. Accordingly, a quadrupling was deduced every two years (consistent with the needs of the system houses) and subsequently modified to ~ 3 years around the mid-later 1970s and currently taken as 3-4 years based on a 1995 assessment [53]. This analysis became enshrined as
Moore’s law and became the productivity criterion by which the IC industry grew at ≈16% compound annual growth rate (CAGR), facilitated by the availability of larger-diameter silicon single crystals to support the requisite larger chip sizes.

The phenomenal growth of the IC industry, achieved by staying on the “productivity learning curve,” continues to be the gauge by which the industry is measured [55,304]. This is evidenced by the cost per bit or logic function historically declining at ≈16% CAGR for the past several decades. This growth has been fueled by four factors; shrinking lithographic design rules, yield improvements, increased equipment utilization and larger wafer diameter. The largest opportunity growth factor to maintain the IC productivity engine and continue on the productivity curve as described by Moore’s law appears to be increased equipment effectiveness; that is, the percentage of time the equipment is adding value to the wafer. The largest challenge to maintaining the productivity curve, however, may be the enormous financial infrastructure required, rather than technological limits to chip density. In that regard, business and manufacturing ideas will become increasingly important to ensure that the long-term productivity growth of the semiconductor industry maintains its growth near historical levels for the next ten years. More than just monitoring productivity, whether by staying on the productivity curve or increasing manufacturing effectiveness, however, is required. Rather, modeling productivity—the identification of new productivity measures—is required [56]. Global specifications, metrology and standards, in addition to CoO opportunities discussed earlier, are important mechanisms to ensure the marketplace reality of the ITRS roadmap trends, based on Moore’s law, is achieved. The cost effectiveness of international standards for emerging technologies such as SOI and 300 mm diameter wafers, in conjunction with the cost-effective production of ICs, such as computer-based design for manufacturability, will offer significant opportunities for an improved quality of life for the world’s citizens. An even greater challenge to maintaining the productivity curve, however, may be the enormous financial infrastructure required, rather than technological limits to chip density [7,305]. This more recent development, regarding the escalating cost of building the IC fabrication facility, has been described, somewhat erroneously, as Moore’s second law [306]. Indeed, a state-of-the-art IC fab producing 65 nm physical gate length MOSFET ICs (130 nm technology generation) is about $2.5B (U.S.). The financial investment for implementation of the 300 mm and 450 mm wafer eras will certainly exacerbate this concern. Accordingly, ingenious engineering advancements to ensure the requisite number of chips per wafer (currently 200 mm in mainstream manufacturing) without introducing the next wafer diameter size, (i.e., 300 mm and, eventually, 450 mm) with the requisite increased chip performance, continues unabated. Nevertheless, the onset of the 300 mm era has begun for those IC manufacturers who anticipate a distinct cost advantage in going to the larger diameter wafer.

Device scaling has been the engine driving this revolution, based in large part on Robert Dennard’s one-transistor memory cell introduced in 1968 [45]. The one-transistor/one-capacitor (1T) cell, in conjunction with the scaling methodology introduced by Dennard et al. in 1974 [46-49] (i.e., reduction in design rules without compromising the current-voltage characteristics) established the paradigm by
which enhanced scaling has progressed and facilitated the explosive growth and applications of the MOSFET IC. The original scaling methodology, based on constant electric field scaling principles, was generalized in 1984 to allow the voltage to be scaled less rapidly than the dimensions by increasing the electric field (with its own scaling factor) [46-49] (see Table 3). Major scaling contributions have occurred in the reductions of gate dielectric thickness, physical gate length and extension junction depth, as discussed by Dennard and colleagues via constant electric-field scaling and, subsequently, constant voltage scaling [47-49]. Dennard has recently summarized several aspects of the generalized scaling relationships [49]:

"The original constant-electric-field scaling principles which we introduced in the early 1970’s were generalized in the 1980’s to allow voltages to be scaled less rapidly than dimensions by increasing the electric field, which has its own scaling factor ε [47-49]. Another generalization can be made by allowing some basic device parameters and the interconnecting wire dimensions to be scaled down by different factors.

With these changes, our current view of scaling is shown in Table 3. Most device physical dimensions are divided by a factor of α_d and the voltage is scaled down by α_d but multiplied by the factor ε, as discussed above. The wiring dimensions and the device width are divided by a factor α_w. A reasonable goal is to increase the circuit speed by a factor α_d, which assumes that the average carrier velocity remains about the same, even if the field ε increases, and that the threshold voltage, V_t, is scaled down by the same factor as the applied voltage, V_DD.

The active power for a given circuit scales as \( \varepsilon^2 / \alpha_d \alpha_w \) while the power density scales as \( \varepsilon^2 / \alpha_w \), assuming that the clock frequency does indeed increase by \( \alpha_d \) and that the same circuit designs are used."

### Table 3: Generalized Scaling Approach [49]. Reproduced by permission of The Electrochemical Society, Inc.

<table>
<thead>
<tr>
<th>Physical Parameter</th>
<th>Generalized Scaling Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel length, L</td>
<td>1/α_d</td>
</tr>
<tr>
<td>Gate insulator, T_OX</td>
<td>1/α_d</td>
</tr>
<tr>
<td>Voltage, V</td>
<td>( \varepsilon / \alpha_d )</td>
</tr>
<tr>
<td>Wiring width</td>
<td>1/α_w</td>
</tr>
<tr>
<td>Channel width, W</td>
<td>1/α_w</td>
</tr>
<tr>
<td>Circuit speed (goal)</td>
<td>α_d</td>
</tr>
<tr>
<td>Circuit power</td>
<td>( \varepsilon^2 / \alpha_d \alpha_w )</td>
</tr>
</tbody>
</table>

**Evolving Directions**

The gate dielectric thickness, physical gate length and extension junction depth parameters were empirically related in 1980 by Simon Sze and colleagues [307] to ensure the retention of long-channel behavior (i.e., no short-channel effects, such as V_t roll-off). For example, consider the scaling of the Kbit DRAM from the early 1970s (4K DRAM) to today’s leadership, high-performance MPU part appropriate for the 90 nm technology generation in 2003 [55]. The SiO_2 gate dielectric has decreased from the range of about 50-100 nm for the 4K DRAM to an anticipated value of about 1.2 nm oxide equivalent thickness (EOT) for the MPU part [55]. Likewise, the physical gate length has decreased from 7.5 µm for the 4K DRAM to a physical gate length of about 45 nm for the MPU part at the 90 nm technology generation [55]. Finally, the junction depth has decreased from several microns for the 4K DRAM to about 20 nm for the extension junction depth for the MPU part at the 90 nm technology generation [55]. In a related fashion, the critical figure of merit for transistor speed, CV/I, has become less than one psec for an NMOSFET and typically approaching one psec (or slightly less) for a PMOSFET, as the MOS physical gate length has decreased from 30 nm to 10 nm, the latter dimension being appropriate for the 64G DRAM and 9 G high-
performance logic era in 2016 [55]. It is anticipated that the “silicon age” of IC microelectronics has a sufficiently robust, but challenging future.

Scaling of the gate dielectric SiO$_2$ to the sub-2 nm regime, however, has exacerbated the occurrence of direct tunneling [308,309] as described by Yuan Taur and colleagues. An extensive global effort is in progress by numerous personnel [310-315] to identify an alternative, high-dielectric constant material to circumvent the gate dielectric direct tunneling leakage current when the silicon oxynitride gate dielectric physical thickness is less than about 1.2 nm for high performance microprocessors, including the relevant diagnostic techniques [8]. Additionally, the importance of gate electrodes (eventually requiring dual metal gates with differing work functions for CMOS optimization or a single, tunable work function metal) and the issues of incorporating the gate stack into an integrated, conventional planar, initially poly electrode, IC process flow has been noted [315]. Al Tasch has clarified the role of the quantum confinement effect in silicon in increasing the effective dielectric thickness of a MOSFET in inversion [316], which cannot be avoided as compared to the poly-depletion effect in the polysilicon gate electrode, which can be negated by utilizing metal gate electrodes. In that regard, the dual metal system with differing work functions is under consideration as the gate electrodes for optimal CMOS performance [55].

Concurrently, a host of studies are in progress to identify an ultrashallow junction fabrication methodology consonant with the sub-90 nm technology generations [55,317]. These studies may be grouped under the classification as classical CMOS structures. On the other hand, a plethora of non-classical CMOS devices are under consideration wherein a unique combination of materials and/or device structural configurations may differ from the conventional or classical planar CMOS structure [55,57]. Of particular importance is the assessment of alternate channels for enhancement of the n- and p-channel mobility, ranging from strained silicon on unstrained Si-Ge on SOI, silicon-germanium on silicon and a host of alternative vertical transistor structural configurations [318,319] as well as the semi-ballistic transistor [320].

Accordingly, the silicon MOSFET’s may be expected to scale in an essentially predictable manner from the present state-of-the-art 90 nm technology generation (MPU physical gate length of 45 nm) to the 22 nm technology generation (MPU physical gate length of 9 nm) [55]. Concurrently, the IC industry has continually seen that the ingenuity of device and process engineers to develop unique device geometries, from the early DRAM era [44] and, more recently, non-classical CMOS devices including vertical transistor configurations or double-gate structures utilizing SOI as appropriate, novel process technologies and models to guide further development may be more influential to IC growth and device performance than might be inferred from the extrapolation of today’s art [57,321,322] and may offer significant unforeseen opportunities. The ubiquitous applicability of Si technology in the information revolution, touching numerous aspects of the lives of the world’s citizens, however, may not necessarily require the state-of-the-art technologies in all applications [7,8]. It is unlikely, moreover, that the present worldwide silicon infrastructure will be regenerated to support a silicon successor [323]. Accordingly, silicon technology is expected to continue as the most powerful driver of the information age for at least the next 100 years, albeit in conjunction with complementary, alternative device structures.

ACKNOWLEDGEMENTS

This manuscript is dedicated to Dr. Gordon K. Teal (1908–2003). The author appreciates discussions and/or critiques of an original version of this manuscript by Kenneth Bean, W. Murray Bullis, Bob Dennard, Jim (James M.) Early, Nick Holonyak, John L. Moll, Walter Runyan, Dieter Schroder, and Frederick Seitz.

REFERENCES


[177] C.J. Frosch and L. Derick, The Oxidation of Silicon to Prevent Surface Erosion During High Temperature Operations, BTL Memorandum #55-113-23 (June 14, 1955)


[184] W.L. Brown, N-Type Surface Conductivity on P-Type Germanium, Phys. Rev., 91, 1365-1376 (1952)


