MOSFET Scaling—The Driver of VLSI Technology

DALE L. CRITCHLOW, FELLOW, IEEE

Invited Paper

This is an introduction to the Classic Paper on MOSFET scaling by R. Dennard et al.*, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," published in the* IEEE Journal of Solid-State Circuits *in October 1974. The history of scaling and its application to very large scale integration (VLSI) MOSFET technology is traced from 1970 to 1998. The role of scaling in the profound improvements in power delay product over the last three decades is analyzed in basic terms.*

Keywords— *CMOS integrated circuits, MOSFET's, MOSFET scaling, scaling, very large scale integration (VLSI).*

I. FOREWORD

The Classic Paper by Dennard *et al.*, "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," [1] published in 1974, is regarded as the seminal reference in scaling theory for MOSFET integrated circuits. (It is referred to as the "scaling paper" in this document.) Integrated circuits experienced exponential improvements in many parameters from the beginning. The "scaling paper" provided the guiding principles to take advantage of these improvements in terms of MOSFET device design, circuit design, and chip design.[1] It is timely to recognize this paper as people are currently attempting to record the phenomenal early history of integrated circuits. Typifying this are the histories recorded by Malone [2] and Bassett [3], which include sources of information not found in the technical journals.

I had the good fortune to be involved in the original work on scaling as well as its application to products spanning almost three decades. My perceptions of the early development and demonstration of scaling MOSFET technology are presented in the first part of this paper. This is followed by my observations of how scaling principles have become dominant factors in strategies for MOSFET integrated circuits.

Manuscript received January 27, 1999; revised January 28, 1999.

The author is with the Electrical and Computer Engineering Department, University of Vermont, Burlington, VT 05405-0156 USA.

Publisher Item Identifier S 0018-9219(99)02989-8.

[1]Gordon Moore published his famous projection in the 1975 *IEDM Digest*, in which he stated that the number of components on a chip essentially doubles each year. This is now popularly known as "Moore's law."

II. INTRODUCTION

Let me begin by reflecting on the situation in 1970, when the work reported in the "scaling paper" began. The period from about 1964 to the early 1970's had experienced an extraordinary revolution in integrated-circuit technology, with many people and organizations making major contributions. A wide array of techniques and approaches were being explored and developed. By 1970, the bipolar junction transistor technologies, which had accumulated over two decades of learning, were being seriously challenged by MOSFET's in applications where chip density and cost factors were of prime importance. Numerous start-up companies, as well as established companies such as Texas Instruments, Fairchild, RCA, Motorola, and companies in Europe and Japan, prepared to exploit the emerging MOSFET technology for logic and memory applications. Some of these parts were already on the market and many were being developed. For example, Intel's first DRAM chip, using a three-device cell [4], was under development and they were developing a microprocessor on a chip.

During this epoch from 1964 to 1970, IBM had also pursued aggressive programs in integrated-circuit logic and memory [5]. A high-level decision was made in January 1968 to abort further development of ferrite and thin magnetic film memory technologies in favor of semiconductor memories. A 128-bit bipolar main-memory chip was qualified in 1969. The company geared up for large-scale manufacturing with volume shipments of high-end machines in early 1971.

The development of MOSFET main memory at IBM lagged a year or two behind the development of bipolar memory. Bob Dennard, Fritz Gaensslen, and I had been part of an IBM Research Division project that developed a high-speed, NMOS, main-memory technology using a six-device cell [5], [6]. This technology was transferred by means of a joint program to the IBM Components Division

in 1967–1968. Operating 512-bit chips were being tested in high-end systems in 1969. By 1970, the qualification and scale-up for large-scale manufacturing of 1- and 2-kbit chips were well underway.

III. Development of the Principles of MOSFET Scaling

In 1970, IBM Research was searching for a technology to fill the cost/performance "file gap" between movable head magnetic disks (which had low cost/bit but high latency time) and random access main memory (which had high performance but high cost/bit) for transaction-based systems. Fixed head files using high-speed drums (or disks) with low latency time were being employed as cache storage. Magnetic bubble technology was emerging as a possible candidate. In mid 1970, Don Rosenheim (Manager of Applied Research) and Sol Triebwasser challenged my department to propose a "monolithic file" for this application. The cost goal was one millicent/bit, a factor of about 1000 times less than the then current cost of main memory.[2] Bob Dennard was the manager of a small group reporting to me. The group included Fritz Gaensslen and Larry Kuhn.

Several types of memory circuits were considered: shift registers; the bucket brigade shift register; charge coupled devices; and the one-transistor DRAM cell. As inventor of the one-transistor DRAM cell [7], Bob was eager to make it a viable candidate and soon proposed a preliminary design, which utilized his cell. Several breakthroughs were required to achieve the technical and cost goals. These included:

- shrinking dimensions on the chip to about 1 μm, which required advances both in lithography and silicon processes;
- dramatic improvements in yield to allow larger chips and higher resolution lithography which, unfortunately, would print smaller, much more numerous defects;
- a means of sensing the very small signals on the bit lines.

During the next several years, each of these problems were solved. The experts in advanced electron beam and optical projection in the Research Division provided leadership for the 1-μm lithography. Hwa Yu and his processing group were the prime movers of advanced processing. They took advantage of the emerging capabilities of ion implantation and dry etching, including basic work on anisotropic Reactive Ion Etch (RIE). The oxide specialists made tremendous strides in learning to grow high-quality, thin gate oxides. Bit-line and word-line redundancy techniques [8], [9] were developed to greatly ameliorate the yield challenges. The latch sense amplifier [10] with the addition of dummy cells [11] solved the sensing problem.

A 5$\times$ shrink of the existing technology was needed to achieve 1-μm dimensions. Bob and I decided that rather than designing the 1-μm technology from scratch, we would scale from some well-characterized devices which had channel lengths of about 5 μm and could be operated with voltages up to 20 V. We observed that if the electric fields were kept constant, the reliability of the scaled devices would not be compromised. In addition, if we could keep the fields in the silicon constant, we would expect fewer problems with short-channel effects and channel-length modulation. Engineers of that era, who relied on slide rule calculations, were well versed in similitude or scaling. Indeed, in the 1960's we had gone through three levels of technology at the research level from 24 V to 12 V to 6 V. We had used rudimentary scaling to guide our device and circuit designs.

A few days later, Bob and Fritz Gaensslen had derived the constant-field scaling theory and its limitations. The scaling theory had remarkable implications on circuit performance, circuit power, and power density, as well as the more obvious chip density. A key to the scaling was that all dimensions including wiring and depletion layers and all voltages including thresholds were scaled in concert. This will be discussed in more detail below. Since that time, the scaling principles have become essential tools in determining strategy and designs for advanced MOSFET integrated-circuit technology.

The "scaling paper" and earlier publications from the same project actually have two very significant contributions:

- the first demonstration of a scaled 1-μm MOSFET suitable for high-speed digital applications;
- the development of the scaling principles and their limitations.

IV. Device Scaling and Hardware Demonstration

Dennard's group immediately dove into demonstrating the viability of MOSFET scaling using two devices:

- an existing, well-characterized 20-V reference device with a 5-μm channel length and a 100-nm gate oxide;
- a scaled 4-V device with 1-μm channel length and a 20-nm gate oxide (the scaled device was similar to the devices actually used in the mid-to-late 1980's with 5-V supplies).

Hwa Yu led the advanced process development effort. The first devices used conventional doping techniques and contact printing. The project was very successful and the essential accomplishments were presented at the International Electron Devices Meeting (IEDM) in 1972 [12]. This was the first formal presentation of the work. At that time, the IEDM only published abstracts, although Dennard recently provided a copy of his slides and text of the presentation [13]. One of the key slides, shown in Fig. 1, summarizes the device scaling principles. All dimensions and voltages, including the threshold voltage, were scaled down by a factor of α while the substrate doping was scaled up by α. (Note that the "scaling paper" uses κ rather than α for the scaling factor.) The structures of the two MOSFET's are shown in Fig. 1 for a scaling factor of five, which should result in scaling down of the threshold voltage and the device current by 5$\times$. In fact, the experimental device characteristics followed the simple scaling laws very

[2] In fact, low-cost DRAM did displace fixed head files by the early 1980's.

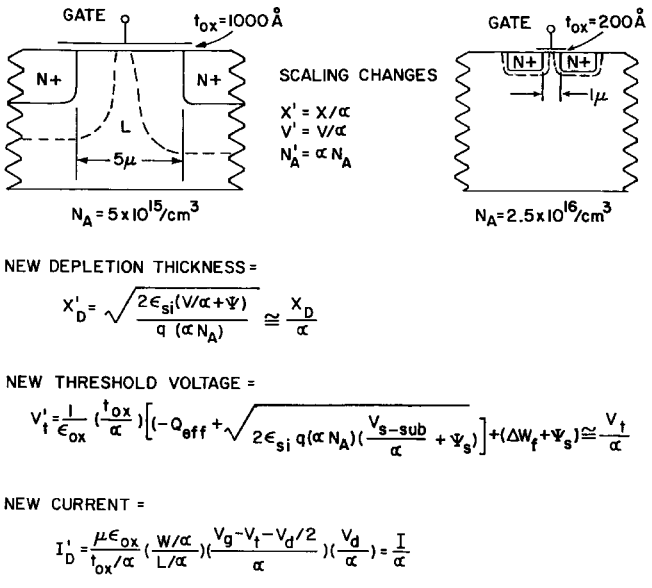


Fig. 1. Device scaling slide from the 1972 IEDM presentation [12].

closely. (The actual experimental curves are given in Figs. 2 and 3 of the "scaling paper.") The IEDM presentation also described the overall circuit and chip scaling principles in some detail.

The final paragraph of the text for the IEDM presentation shows that the authors were well aware of the potential for scaling to achieve high-speed MOSFET circuits.

> Finally, we conclude that the performance gains to be achieved with miniaturization of MOS devices, particularly when combined with other improvements in processing and structures which appear to be in the making, will be truly phenomenal.

Dennard wrote an invited paper [14] in 1993, in which he records his perceptions of how the paper was received by the IEDM audience.

> When I presented the talk at the IEDM in Washington, the attendees expressed a lot of interest, but also a great deal of disbelief. There had been only a little previous work on small dimension MOS transistors aimed at high-frequency amplifiers, and ours was the first 1-μ m transistor fully characterized and shown to be suitable for digital circuit operation. However, when I talked about the need for a 200-Å^0 insulator in our device, I heard a lot of laughter in the audience. At that time, many people thought that it was difficult to make thin insulators and that 1000-Å^0 was near the limit of practical use!

A small memory array, which used the scaled devices, was built utilizing electron beam exposure in cooperation with T. H. P. Chang and his advanced electron beam lithography group. A paper was published in 1973 [15].

Once the basic device scaling principles were demonstrated, Dennard, Yu, and their groups moved to take full advantage of the newly emerging ion-implantation capabilities to tailor the substrate doping for optimum turn-on characteristics and to build self-aligning gates with

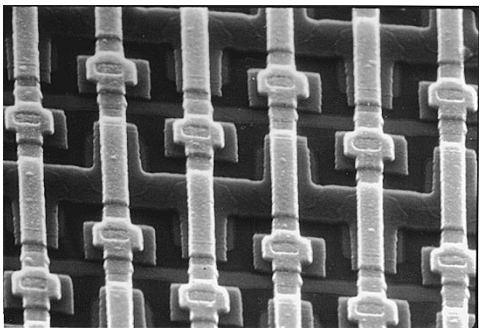


Fig. 2. SEM photo of cells on 8-kbit DRAM chip using 1.25-μm lithography published in 1975 [19].

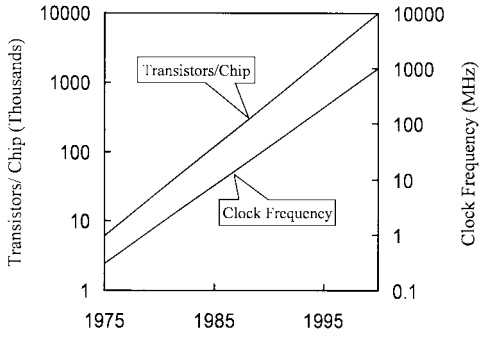


Fig. 3. Trends of chip density and clock frequency for leading edge microprocessor products over two and one half decades.

shallow junctions. Further improvements to the device designs were made with Andre LeBlanc from the IBM Burlington Laboratory and Leo Rideout joining the effort. Finally, the "scaling paper" [1] was published in 1974.

Other researchers were also investigating the design of very small devices and low operating voltages. For example, Hoeneisen and Mead published an excellent theoretical paper [16] in 1972 that dealt with some of the fundamental limitations of MOSFET's. They projected that it would be possible to build a 2-V transistor with a channel length of 0.4 μm and a 14-nm gate oxide. That same year, Swanson and Meindl [17] described very low-power, low-speed CMOS circuits which could operate with voltages as low as 0.2 V.

V. Demonstration of Scaling of an 8-kbit DRAM

Once the scaled device design was demonstrated, Dennard and Yu led an effort to build a complex scaled DRAM chip using electron beam exposure. They started with the design data from an 8-kbit PMOS DRAM developed at the IBM Burlington Laboratory [18]. The dimensions were scaled from 3.75 to 1.25 μm. It was found that the available etching techniques were not satisfactory. Single, isolated devices with dimensions of 1.25 μm could be built using isotropic (wet or dry) etching available. However, the processes were not adequate for the complex, tightly packed structures on a DRAM chip. Hwa Yu developed an anisotropic dry etching process providing the breakthrough required. This was perhaps the first use of dry anisotropic etching for complex structures at these dimensions. The

Table 1
Idealized Scaling

Parameter	Constant Field Scaling	Constant Voltage Scaling
Dimensions	$1/\kappa$	$1/\kappa$
V_{DD}	$1/\kappa$	1
Fields	1	κ
V_T	$1/\kappa$	1
Current	$1/\kappa$	κ
Capacitances	$1/\kappa$	$1/\kappa$
Delay Time	$1/\kappa$	$1/\kappa^2$
Power/Circuit	$1/\kappa^2$	κ
Power x Delay	$1/\kappa^3$	$1/\kappa$
Power/Area	1	$1/\kappa^3$
Line Resistance	κ	κ
RC	1	1
IR/V_{DD}	κ	κ^2

scaled 8-kbit chip was successfully implemented and a paper was published in 1975 [19]. The SEM photo in Fig. 2, which is from that paper, shows the remarkable delineation achieved. This demonstrated scaling in a dramatic way and had a major impact on IBM and many people in the industry.

Our attention was then focused on using scaling to develop a 1-μm high-speed MOSFET technology for high-performance SRAM and logic chips, one of the goals being to replace bipolar transistors in mainframe computers. This work included an investigation of operation at liquid nitrogen temperature to take advantage of the higher mobility and the sharper turn-on characteristics. A series of eight papers [20] that took full advantage of the scaling principles was published in 1979.

VI. Constant-Field Scaling Versus Constant-Voltage Scaling

The essence of constant-field scaling is summarized in Table 1, in which ideal scaling is assumed and the second-order effects are neglected. As noted above, all dimensions (including wiring) and voltages of a circuit are scaled in concert by a factor of κ. The doping level of the substrate is increased by κ so that the depletion layer thickness scales down with κ. The circuit gets faster by κ, the power/circuit is reduced by κ^2, the power delay product improves by κ^3, and the power/unit area remains constant.

These remarkable results demonstrated that we could move to very high levels of integration while making the chips faster and keeping the power dissipation at reasonable levels. For example, scaling by $5\times$ provides $25\times$ more circuits with the same chip size and results in $5\times$ performance increase with no increase in chip power. Of course, as the number of circuits/chip is increased, the space required for interconnections on the chip tends to increase. This can counteract the density, performance, and power improvements somewhat. However, in practice, putting more function on a chip has efficiencies, which offset the effects of the extra wiring requirements. In addition, extra levels of wiring can be used.

The results are quite different when all of the dimensions, but not the voltages, are scaled. This is referred to in Table 1 as constant-voltage scaling. The circuits are faster by κ^2 rather than κ as was the case with constant-field scaling. However, this neglects the fact that velocity saturation and intrinsic device resistances limit the performance gain, particularly at very small device dimensions. Note that the power/circuit increases by κ and the power delay product improves only by $1/\kappa$. Perhaps most important is that the power-per-unit area increases by κ^3 rather than being constant as in constant-field scaling. Since the fields in the gate oxide and silicon increase by κ, critical questions arise regarding gate oxide reliability, velocity saturation, and hot electron damage. In addition, the sharpness of the MOSFET turn-on characteristics and the variation of threshold voltage with channel-length and applied voltages become more problematical.

In both types of scaling, the interconnection resistance effects become more pronounced as dimensions are shrunk as shown in Table 1. The $R_{\mathrm{Wire}}C_{\mathrm{Wire}}$ of the wiring stays constant while the circuits become faster. The IR_{Wire} drops in the ground and power supply lines become larger fractions of the power supply and threshold voltages. Solutions to these limitations are discussed in Section IX.

VII. The 5-V Era

Most of the early MOSFET products used PMOS devices with power supply voltages ranging from 10 to 20 V with saturated or linear load devices. There were soon tremendous industry pressures to insure compatibility with 5-V TTL chips. Therefore, MOSFET chips moved rather quickly in the early 1970's to TTL standards, first by providing a TTL interface and, soon after, moving to a single 5-V supply. In many cases, reducing the power supply voltages to 5 V had the effect of actually reducing the fields substantially with rather significant performance losses, unless the devices were scaled accordingly. In the early 1970's, most companies using PMOS converted to NMOS to achieve higher performance. The NMOS depletion load device [21] was introduced to allow full rail-to-rail voltage swings and improvements in power and performance. Device capacitances due to gate overlap and deep source/drain diffusions were reduced dramatically from those of the early devices.

The technologies were shrunk repeatedly (both in the vertical direction as well as in the plane of the wafer) over a period of at least 15 years with the supply voltage kept at 5 V. Improvements in materials, processing, and transistor design allowed operation at higher electric fields and the circuit speeds improved over the years. As noted above, because of velocity saturation and intrinsic device resistance effects, the performance gains of actual devices were much less than those predicted by ideal constant-voltage scaling. As the circuits were scaled, the power/circuit increased dramatically, which ultimately resulted in cooling limitations. In addition, the high fields raised reliability concerns.

The possibility of reducing the power supply voltages to 3.3 V to alleviate the power problem started to be pondered seriously in the industry by the mid-1980's. There were many discussions at technical conferences and in the standards committees and several companies considered using developing 3.3-V technologies. This was met with high levels of resistance from systems and marketing people since the industry had a tremendous investment in 5-V parts. In addition, interfacing between 5 and 3.3-V parts was problematic. However, there was another, less obvious, factor. The tolerances on channel length must scale with channel length in order to maintain an acceptable part-to-part performance spread. For a given performance, a device designed for high voltage operation will have a longer channel than one designed for lower voltage operation. Therefore, for a given lithography tolerance, the higher voltage device has a lower percentage tolerance in channel length resulting in lower part-to-part spreads in performance.

CMOS provided a solution to the power problem since it essentially eliminated the static or dc power. This extended the useful lifetime of 5-V by several years. As additional functions were integrated on the chip, the fraction of the time during which a given CMOS circuit was active was greatly reduced (on the average). This also helped maintain reasonable chip power.

The next barriers to overcome for 5-V devices were high field effects, e.g., hot electrons, which limited channel lengths to about 1–1.5 μm. Improved device structures, such as graded junctions and the lightly doped drain (LDD) MOSFET using a spacer technology [22], provided relief and extended the 5-V technologies to channel lengths of about 0.6 μm.

By the late 1980's, there were two intense drives in the industry:

- very high-performance systems such as workstations;
- portable, battery-powered equipment requiring much lower power, but also requiring high performance; low-voltage, low-performance, battery-powered applications such as wristwatches and portable electronic devices had been important markets for many years.

It was increasingly clear that power dissipation and reliability were rapidly becoming limiting factors for high performance 5-V systems. Finally, it was necessary to scale to lower voltages for products in the 1990's and beyond.

VIII. Scaled Power Supplies—After 5 V

By the late 1980's, the industry was moving quickly to lower voltages for leading edge applications in high-speed processors. Technologies optimized at 3.3 V [23] and high-speed 3.3-V processor chips [24], [25] were being developed. Lower voltage DRAM chips for portable equipment were under development [26]. The operating voltages of leading edge products have moved rapidly from 3.3 V to 2.5 V [27], [28] to 1.8 V and lower [29]. Several 1.8-V chips (presently in early development and manufacturing, with channel lengths of less than 0.15 μm and gate oxides of about 4 nm) have been described [30]–[32]. Clock rates up to 1 GHz have been demonstrated on experimental 64-bit microprocessors [33]. The 1997 SIA *National Technology Roadmap* [34] projects power supply voltages of about 1.2 V and gate oxide thickness of about 2–3 nm by 2003. A number of excellent papers have been written over the last few years which describe the advances in low-power and high-performance systems [35]–[41].

IX. Dealing with the Scaling Limitations

As described in the "scaling paper," the basic limitations of scaling are the threshold characteristics of the MOSFET and resistance effects for the interconnections. The magnitude of the threshold voltage does not scale well for voltages less than about 1.5 V. In addition, short-channel effects, where the threshold varies as the channel length and applied voltages change, cause degradation of circuit performance. Device designers have succeeded in improving the turn-on characteristics of the devices by optimizing the doping profiles of the channel region and the source-drain electrodes using ion implantation, as described in the "scaling paper." The move to CMOS has made the circuits less sensitive to body, or substrate, effects, allowing higher doping to be used. More sophisitcated approaches to scaling, sometimes referred to as "generalized scaling," have been developed [42]–[44]. This approach essentially utilizes a compromise between constant-field scaling and constant-voltage scaling in which the gate oxide thickness and the channel length scale more rapidly than the supply voltage.

Circuit designers have learned to design high-speed circuits with larger off-currents and larger short-channel effects. In addition, multiple device designs, for example, the use of different gate oxide thicknesses, each tailored to a specific purpose, can be used on the same chip. SOI shows promise for achieving improved characteristics. Circuit designers are now developing circuits that are more tolerant of the short-channel effects. The net result is that it is now feasible to scale the threshold voltage lower than what was deemed practical even a few years ago, keeping the door open for further scaling of the technology in the future.

The resistance problem has been dealt with by a number of techniques.

- Wires have not been shrunk in the vertical direction as much as scaling rules prescribe. This allows a tradeoff between the wiring resistances and capacitances.
- Silicides have been added to polysilicon gates and diffusions to lower the electrode resistances.
- Additional levels of metal have been added to contain the $R_{\mathrm{Wire}}C_{\mathrm{Wire}}$ problems by allowing wider, thicker wiring at the higher wiring levels. Similarly, very thick, wide wires are used at the upper levels to provide adequate ground and power distribution systems.
- Copper wiring has been introduced recently to reduce resistance.
- Design tools have been developed which allow optimization of wiring to minimize the effects of resistance.

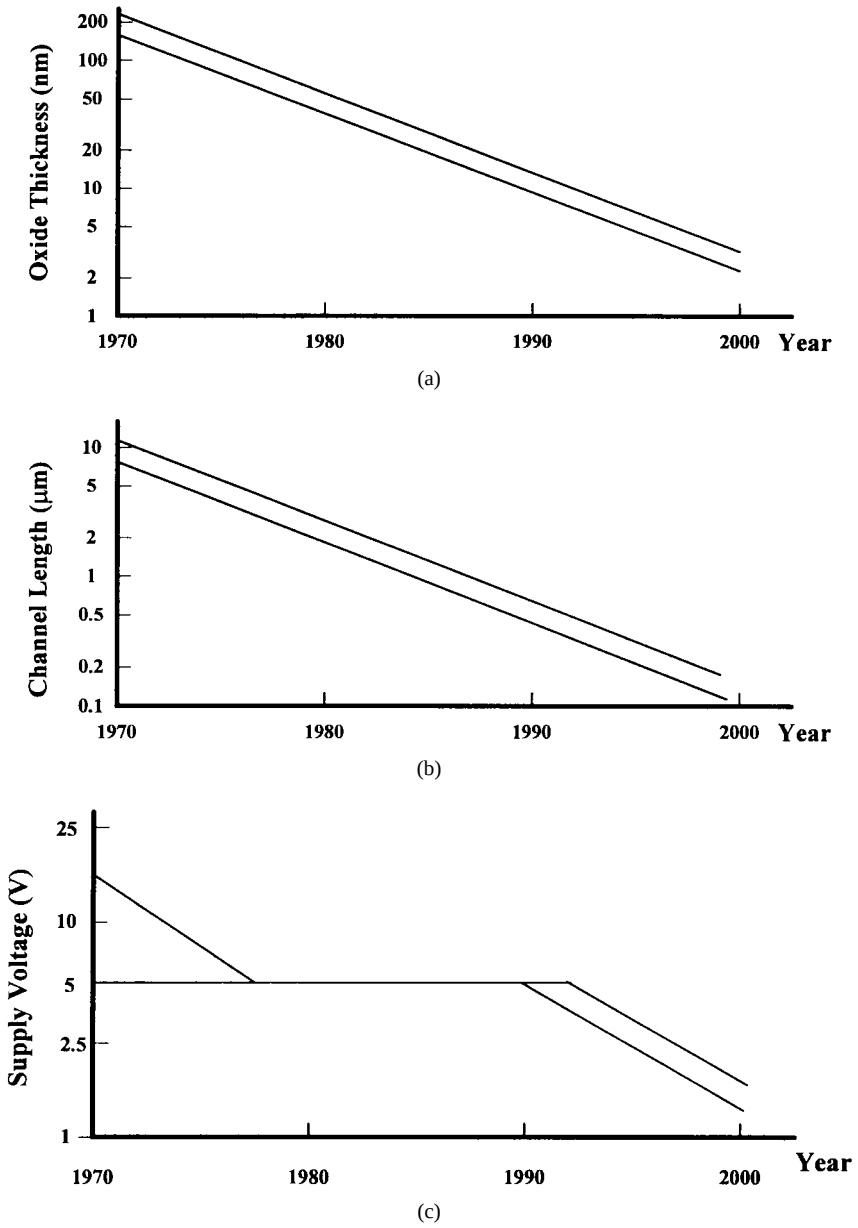


Fig. 4. Long-term MOSFET technology trends showing: (a) gate oxide thickness; (b) channel length; and (c) power supply voltage.

A good overview of the technology and design of the wiring for leading edge high-performance chips using 0.22-μm CMOS technology with copper wiring is given in [45]. The many tradeoffs among chip density, wire length, and performance have been treated rather thoroughly [46], [47].

X. The Bipolar Challenge

From the early days of scaling, it was projected that at some point MOSFET technology would replace bipolar technology for high-speed machines. The relative simplicity, high packing density, low power, and ease of scaling of MOSFET's were essential leverages. The net result was that much more functionality could be built on a MOSFET chip with the attendant advantages. This led to CMOS performance similar to that of bipolar technology but at a much lower product cost. Finally, by the early 1990's, CMOS was becoming the dominant high-end, high-speed digital technology [25], [48]–[49].

XI. An Overview of Scaling History

The plots of Fig. 3 illustrate the profound advances in chip density [50] and clock frequency [47] over 25 years. The number of transistors/chip has increased by about 1600× (1.35×/year). The clock frequency has increased by about 3000× (1.38×/year). It should be noted that while clock frequency is a very convenient and common measure of system performance, it is not particularly accurate since the actual computing power of the system is influenced strongly by the particulars of the system architecture.

It is interesting to reflect on the history of scaling over three decades. The overall trends for device parameters in leading edge microprocessors are shown in Fig. 4. (A spread of 1.5× is shown to account for variations among manufacturers.) Initially, the device technologies and designs were relatively crude and there were wide variations in the power supply voltages. The need to be TTL compatible led to the widespread adoption of 5-V supplies by the mid-1970's. As mentioned heretofore, many of the MOSFET technologies in the late 1970's could have been used an operating voltage exceeding 5 V (with the attendant higher fields). This would have given higher performance. Therefore, the move to 5-V MOSFET chips was a compromise in many cases, especially if power was not a limitation.

During the period from about the mid-1970's to the late 1980's, essentially constant-voltage scaling prevailed. Continuous improvements in gate oxide processing allowed increasingly larger electric fields to be used. CMOS was introduced and improved device designs allowed shorter and shorter channel lengths. Finally, in about 1990, chip power limitations, coupled with reliability concerns, propelled the movement to lower supply voltages.

From about 1990 and into the next decade, the strategy was that of reducing the voltage about 1.5× every few years. The scaling, while similar to the scaling principles put forth in the "scaling paper," are much more sophisticated, and "generalized-scaling" theories have been developed.

Determining and accurate history of device performance over three decades is difficult since much of the data in early papers were incomplete. Therefore, for this paper, the overall performance gain from 1975 to 1998 was estimated using the simple metric CV_{DD}/I. The variable C is the total gate capacitance, V_{DD} is the power supply voltage, and I is the device current with both gate and drain tied to V_{DD}. Experimental data from published papers [51], [52] gives a rate of increase of basic device performance of approximately 1.2×/year from 1970 to 2000. This corresponds to a performance increase of approximately 100× from 1975 to 2000. (This is less than $V_{DD}/L_{\mathrm{EFF}}^2$ would suggest. The difference is due to velocity saturation and resistance effects in real devices.) The additional 30×, or so, needed to explain the clock frequency trend curve of Fig. 3 is due to a host of improvements including improved circuits, CMOS, the leverages afforded by including more functionality on a chip, improved architectures, and improved design methodologies.

At some point, the limitations of silicon technology will result in diminishing returns and other strategies will dominate. When this will happen is not clear. Historically, people have found ways to remove, or work around, limitations in silicon technology. SOI, which allows fully depleted devices, is in its early stages. Low-temperature operation may become a viable option. Circuit designers are inventing new modes of operation to counteract short-channel and threshold effects.

XII. Scaling as an Educational Tool

The use of principles of scaling is particularly useful for teaching electronics at the undergraduate and graduate levels and has found its way into textbooks [53]. It allows the development of very clear, straightforward analyses to quantify the tradeoffs which drive integrated circuit technology. The symbiotic relationships among device structures, device characteristics, and circuit performance and power become very clear to the students.

XIII. Conclusions

Scaling has been a fundamental driving force in MOSFET circuits for several decades. Although the concepts underlying scaling are rather simple, the implications and results are profound. The "scaling paper" has served as the basic reference to the industry for almost two and a half decades and promises to be useful for another. As the MOSFET technology has matured and more limitations must be considered, the scaling approaches have become much more sophisticated. Many people throughout the industry have contributed to advances in scaled MOSFET technologies and designs over the last several decades. Several of the authors of the "scaling paper" have dedicated their careers to implementing and improving on the scaling principles in advanced chips and products with a series of papers over many years.

Acknowledgment

The help given by Dr. J. D. Meindl (Georgia Institute of Technology), Dr. L. M. Terman, Dr. E. J. Nowak, Dr. R. H. Dennard (IBM), and Dr. C. Alajajian (University of Vermont) are greatly appreciated. They reviewed the paper and provided excellent suggestions and information.

References

[1] R. H. Dennard, F. H. Gaensslen, H. 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*, vol. SC-9, pp. 256–268, Oct. 1974; see also this issue, pp. 668–678.

[2] M. Malone, *Microprocessors—A Biography*, *Book*. Santa Clara, CA: Springer-Verlag, 1995.

[3] R. Bassett, "New technology, new people, new organizations: The rise of the MOS transistor, 1945–1975," Ph.D. dissertation, Princeton University, Princeton, NJ, Jan. 1998.

[4] W. Regitz and J. Karp, "Three-transistor-cell 1024-bit 500-ns MOS RAM," *IEEE J. Solid-State Circuits*, vol. SC-5, pp. 181–186, Oct. 1970.

[5] E. Pugh, D. Critchlow, R. Henle, and L. Russell, "Solid state memory development in IBM," *IBM J. Res. Develop.*, vol. 25, pp. 585–602, Sept. 1981.

[6] P. Pleshko and L. Terman, "An investigation of potential MOS transistor memories," *IEEE Trans. Electron. Comput.*, vol. EC-15, pp. 423–427, Aug. 1966.

[7] R. Dennard, "Field-effect transistor memory," U.S. Patent 3 387 286, June 4, 1968.

[8] S. E. Schuster, "Multiple word/bit line redundancy for semiconductor memories," IBM Internal Rep. RC 3410, June 17, 1971.

[9] ____, "Multiple word/bit line redundancy for semiconductor memories," *IEEE J. Solid-State Circuits*, vol. SC-13, p. 698, Oct. 1978.

[10] D. L. Critchlow, "Sense amplifer for IGFET memory," *IBM Tech. Disclosure Bulletin*, vol. 13, no. 6, p. 1720, 1970.

[11] K. Stein, A. Sihling, and E. Doering, "Storage array and sense/refresh circuit for single-transistor memory cells," in *Dig. Tech. Papers ISSCC*, Feb. 1972, pp. 56–57.

[12] R. Dennard, F. Gaensslen, L. Kuhn, and H. Yu, "Design of micron MOS switching devices," presented at IEEE International Electron Devices Meeting (IEDM), Dec. 1972.

[13] ——, "Design of micron MOS switching devices," presented at IEDM, Dec. 1972, notes used in presentation.

[14] R. Dennard, "The starting point of an idea—The MOS scaling rule" (in Japanese), *Nikkei Microdevices No. 91*, pp. 139–40, Jan. 1993.

[15] H. Yu, R. Dennard, T. H. P. Chang, and M. Hatzakis, "An experimental high-density memory array fabricated with electron beam," in *Dig. Tech. Papers ISSCC*, Feb. 1973, pp. 98–99.

[16] B. Hoeneisen and C. Mead, "Fundamental limitations in microelectronics—I. MOS technology," *Solid State Electron.*, vol. 15, no. 7, pp. 819–829, July 1972.

[17] R. Swanson and J. Meindl, "Ion-implanted complementary MOS transistors in low-voltage circuits," *IEEE J. Solid-State Circuits*, vol. SC-7, pp. 146–153, Apr. 1972.

[18] W. Hoffman and H. Kalter, "An 8 Kb random-access memory chip using the one-device FET cell," *IEEE J. Solid-State Circuits*, vol. SC-8, pp. 298–305, Oct. 1973.

[19] H. Yu, R. Dennard, T. H. P. Chang, C. Osburn, V. DiLonardo, and H. Luhn, "Fabrication of a miniature 8-Kbit memory chip using electron-beam exposure," *J. Vac. Sci. Technol.*, vol. 12, no. 6, p. 1297, Nov./Dec. 1975.

[20] "1 μm MOSFET VLSI technology: Parts I–VIII," *IEEE J. Solid-State Circuits*, vol. SC-14, pp. 240–301, Apr. 1979.

[21] T. Masuhara, M. Nagata, and N. Hashimoto, "A high performance N-channel MOSLSI using depletion-type load elements," *IEEE J. Solid-State Circuits*, vol. SC-7, pp. 224–231, June 1972.

[22] P. Tsang, S. Ogura, W. Walker, J. Shepard, and D. Critchlow, "Fabrication of high-performance LDDFET's with oxide sidewall-spacer technology," *IEEE Trans. Electron Devices*, vol. ED-29, pp. 590–596, Apr. 1982.

[23] A. Bhattacharyya, R. Mann, E. Nowak, R. Piro, J. Springer, S. Springer, and D. Wong, "A half-micron manufacturable high performance CMOS technology applicable for multiple power supply applications," in *Proc. Int. Symp. VLSI Technology, Systems and Applications*, Taipei, Taiwan, May 1989, pp. 321–326.

[24] R. Allmon, W. Bowhill, B. Benschneider, J. Brown, E. Cooper, W. H. Durdan, A. Fisher, M. Gavrielov, P. Gronowski, W. Grundmann, W. Herrick, D. Kravitz, W. Maheshwari, R. Marcello, G. Mills, M. Mittal, V. Peng, J. Pickholtz, S. Samudrala, D. Sanders, R. Stamm, P. Starvaski, and W. Wheeler, "CMOS implementation of a 32b computer," in *Dig. Tech. Papers ISSCC*, Feb. 1989, pp. 80–81.

[25] H. Schettler, W. Haug, K. Getzlaff, C. Starke, and A. Bhattacharyya, "A mainframe processor in CMOS technology with 0.5 μm channel length," *IEEE J. Solid-State Circuits*, vol. 25, pp. 1166-1177, Oct. 1990.

[26] M. Aoki, J. Etoh, K. Itoh, S. Kimura, and Y. Kawamoto, "A 1.5-V DRAM for battery-based applications," in *Dig. Technical Papers ISSCC*, Feb. 1989, pp. 238–9.

[27] B. Davari, W. Chang, M. Wordeman, C. Oh, Y. Taur, K. Petrillo, D. Moy, J. Bucchignano, H. Ng, M. Rosenfield, F. Hohn, and M. Rodriguez, "A high performance 0.25 μm CMOS technology," in *Dig. IEDM*, Dec. 1988, pp. 56–59.

[28] H. Sanchez, L. Eisen, C. Croxton, A. Piejko, C. Nicoletta, I. Vo, B. Branson, W. Wang, Q. Nguyen, T. Buti, L. Hsu, M. Saccamango, S. Ratanaphanyara, R. Philip, J. Alvarez, S. Weitzel, and G. Gerosa, "A 200-MHz 2.5-V 4 W superscalar RISC microprocessor," in *Dig. Tech. Papers ISSCC*, Feb. 1996, pp. 218–219.

[29] Y. Taur, Y. Mii, D. Frank, H. Wong, D. Buchanan, S. Wind, S. Rishton, G. Sai-Halasz, and E. Nowak, "CMOS scaling into the 21st century: 0.1-μm and beyond," *IBM J. Res. Develop.*, vol. 39, no. 1–2, pp. 245–260, Jan./Mar. 1995.

[30] J. Schutz and R. Wallace, "A 450 MHz IA32 P6 family microprocessor," in *Dig. Tech. Papers ISSCC*, Feb. 1998, pp. 236–7.

[31] H. Kubosawa, H. Takahashi, and S. Mitarai, "A 1.2-W, 2.16-GOPS/720-MFLOFS embedded superscalar microprocessor for multimedia applications," *IEEE J. Solid-State Circuits*, vol. 33, pp. 1640–1648, Nov. 1998.

[32] C. Akrout, J. Bialas, M. Canada, D. Cawthron, J. Corr, B. Davari, R. Floyd, S. Geissler, R. Goldblatt, R. Houle, P. Kartschoke, D. Kramer, P. McCormick, N. Rohrer, G. Salem, R. Schulz, L. Su, and L. Whitney, "A 480-MHz RISC microprocessor in a 0.12-μm L_{eff} CMOS technology with copper interconnects," *IEEE J. Solid-State Circuits*, vol. 33, pp. 1609–1616, Nov. 1998.

[33] J. Silberman, N. Aoki, D. Boerstler, J. Burns, S. Dhong, A. Essbaum, U. Ghoshal, D. Heidel, P. Hofstee, K. Lee, D. Meltzer, H. Ngo, K. Nowka, S. Posluszny, O. Takahashi, I. Vo, and B. Zoric, "A 1.0-GHz single-issue 64-bit powerPC integer processor," *IEEE J. Solid State Circuits*, vol. 33, pp. 1600–1608, Nov. 1998.

[34] Semiconductor Industry Association, *The National Technology Roadmap for Semiconductors*, 1997.

[35] K. Itoh, "Trends in megabit DRAM circuit design," *IEEE J. Solid-State Circuits*, vol. 25, pp. 778–789, June 1990.

[36] K. Itoh, K. Sasaki, and Y. Nakagome, "Trends in low-power RAM circuit technologies," *Proc. IEEE*, vol. 83, pp. 524–543, Apr. 1995.

[37] D. Singh, J. Rabaey, M. Pedram, F. Catthoor, S. Rajyopal, N. Sehgal, and T. Mozdzen, "Power conscious CAD tools and methodologies: A perspective," *Proc. IEEE*, vol. 83, pp. 570–5594, Apr. 1995.

[38] B. Davari, R. Dennard, and G. Shahidi, "CMOS scaling for high performance and low power—The next ten years," *Proc. IEEE*, vol. 83, pp. 595–606, Apr. 1995.

[39] H. Stork, "Technology leverage for ultra-low power information systems," *Proc. IEEE*, vol. 83, pp. 607–618, Apr. 1995.

[40] J. Meindl, "Low-power microelectronics: Retrospective and prospect," *Proc. IEEE*, vol. 83, pp. 619–635, Apr. 1995.

[41] E. Harris, S. Depp, W. Pence, S. Kirkpatrick, M. Sri-Jayantha, and R. Troutman, "Technology directions for portable computers," *Proc. IEEE*, vol. 83, pp. 636–658, Apr. 1995.

[42] P. Chatterjee, W. Hunter, T. Holloway, and T. Lin, "The impact of scaling laws on the choice of n-channel or p-channel for MOS VLSI," *IEEE Electron Device Lett.*, vol. EDL-1, pp. 220–223, Oct. 1982.

[43] G. Baccarani, M. Wordeman, and R. Dennard, "Generalized scaling theory and its application to a 1/4 micron MOSFET design," *IEEE Trans. Electron Devices*, vol. 31, pp. 452–62, 1984.

[44] R. Dennard, "Power supply considerations for future scale CMOS systems," in *Proc. Int. Symp. VLSI Technology, Systems and Applications*, Taipei, Taiwan, May 1989, pp. 188–192.

[45] A. Stamper, "Interconnection scaling to 1 GHz and beyond," *IBM MicroNews*, vol. 2, no. 4, pp. 1–7, 1998.

[46] J. Davis, V. De, and J. Meindl, "A stochastic wire-length distribution for gigascale integration (GSI), Parts I and II," *IEEE Trans. Electron Devices*, vol. 45, pp. 580–597, Mar. 1998.

[47] D. Ferry and L. Akers, "Scaling theory in modern VLSI: Factors affecting interconnects, wire length and clock speed," *IEEE Circuits and Devices*, vol. 13, pp. 41–44, Sept. 1997.

[48] A. Masaki, "Deep-submicron CMOS warms up to high-speed logic," *IEEE Circuits and Devices*, Nov. 1992, pp. 18–24.

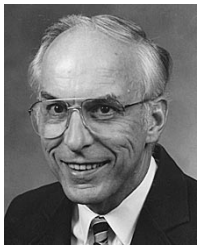
[49] G. Sai-Halasz, "Performance trends in high-end processors," *Proc. IEEE*, vol. 83, p. 20, Jan. 1995.

[50] W. Lattin, J. Bayliss, D. Budde, S. Colley, G. Cox, A. Goodman, J. Rattner, W. Richardson, and R. Swanson, "VLSI microprocessor systems," in *Dig. Technical Papers ISSCC*, Feb. 1981, pp. 110–111.

[51] D. Critchlow, R. Dennard, and S. Schuster, "Design of large-scale integrated logic circuits using MOS devices," in *Proc. 6th Annu. Microelectronics Symp.*, St. Louis, MO, June 1967, pp. C5-1–C5-8.

[52] M. Bohr, S. S. Ahmed, S. U. Ahmed, M. Bost, T. Ghani, J. Greason, R. Hainsey, C. Jan, P. Packan, S. Sivakumar, S. Thompson, J. Tsai, and S. Yang, "A high performance 0.25 μm logic technology optimized ro 1.8 V operation," in *Dig. IEDM*, 1996, pp. 847–850.

[53] C. Mead and L. Conway, *Introduction to VLSI Systems*. Reading, MA; Addison-Wesley, 1980, pp. 33-37



Dale L. Critchlow (Fellow, IEEE) received the B.S. degree in electrical engineering from Grove City College, Grove City, PA, in 1953 and the M.S. and Ph.D. degrees in electrical engineering from Carnegie Institute of Technology (CIT), Pittsburgh, PA, in 1954 and 1956.

He was Assistant Professor of Electrical Engineering at CIT from 1956 to 1958. He joined IBM Research in 1958, where he worked on tunnel diode circuits and digital data transmission techniques until 1964. He joined the MOSFET development in the IBM Research Laboratory in 1964 and managed the Device and Circuit Design Department during early development of the NMOS MOSFET technology. He managed the Silicon Engineering Area in Research from 1970 to 1976, which developed 1-μm MOSFET devices and circuits for DRAM, logic, and SRAM. He managed an advanced development group in the Components Division in East Fishkill, NY, from 1977 to 1981. During this period, his group developed the LDD technology using oxide spacers. He transferred to the Microelectronics Division of IBM in Essex Junction, VT, in 1981 where he directed advanced development work on all generations of DRAM technology from 1–256-Mc chips. Upon retiring in 1993, he joined the Electrical and Computer Engineering Department, University of Vermont, Burlington, in 1993.

Dr. Critchlow was an IBM Fellow and is a member of the National Academy of Engineers.