Recollections on MOSFET Scaling

By Dale L. Critchlow, IBM Fellow, Retired; Dale.Critchlow@uvm.edu

The Beginnings

By 1970 the MOSFET technology was finding its way into manufacturing in a number of companies.[1] Bob Dennard and I were part of the team that developed the NMOS technology (2) in the T. J. Watson IBM Research Laboratories in the 1960s. The first IBM NMOS MOSFET product, which was entering large scale manufacturing, was a high-speed main-memory with a 50ns typical access time (100ns spec) at the board level. It used 1Kb chips (soon replaced with 2Kb chips) with a six-device cell using off-chip bipolar sense amplifiers and high level decoders proposed by Peter Pleshko and Lewis Terman (3). These chips replaced the bipolar main-memory technology which had been introduced a couple of years earlier to replace ferrite core memory.

In mid-1970, IBM Research management was searching for a technology to fill the "file gap" between moveable head magnetic disks and random access main-memory for transaction based systems. This performance gap was being filled by expensive fixed head HDDs which had much smaller latency time than the moveable head HDDs. Don Rosenheim (Manager of Applied Research) and Sol Triebwasser challenged my department to develop a proposal for a "monolithic file" with a cost/bit of about 1 millicent/bit or 1/1000 of the projected main-memory cost. Bob Dennard was manager of a small group including Fritz Gaensslen and Larry Kuhn which reported to me. There were a number of options including shift registers and CCDs, but Dennard as the inventor was keen on pursing the one-transistor DRAM cell. Bob did some preliminary analyses, and concluded that we would need feature sizes of about 1μm, a 5X shrink from those in manufacturing, to achieve our goals.

We realized that we would have to scale the vertical dimensions (oxide thickness and junction depth) and adjust the doping level of the substrate to maintain usable device characteristics. Further, we would have to scale the operating voltages as well to preserve reliability and limit power dissipation. In fact, we had done this twice before in the 1960s, first from 24V to 12V and then to 6V using rudimentary scaling to guide our designs. (Engineers of that era, before the advent of computer simulation, were well versed in design by similitude or scaling.) We observed that our current transistors with channel lengths of 5μm and gate oxide thickness of 100nm could be operated at 20V. Therefore, we could scale to a 4V power supply with a 1um, 20nm transistor. We noted that the circuits would consume less power and be faster. Within a few days Bob, Fritz and Larry had formalized the constant-field scaling theory and its limitations.

The implications of scaling were remarkable. If all dimensions, voltages (including threshold voltage) and doping levels were scaled by a constant factor κ: a) the circuit delay was decreased by κ, b) the power/circuit was decreased by κ^2, and c) the power delay product was reduced by κ^3. Further, the power/unit area of silicon remained constant! These were exactly the results we needed to develop a competitive low cost memory. On the down side, there were questions about the scalability of the threshold voltage and the fact that the IR drops and RC time constants of the interconnects become more severe with scaling. Of course, there were a host of transistor design, process and reliability challenges.

At that point, we were convinced that MOSFET memory would replace fixed head files. Further, we speculated that it may also replace moveable head disk storage for some applications. We also started to believe that the MOSFET would someday replace the bipolar transistor in high-performance logic and memory applications.

Driving the Demonstration and Implementation of Scaling was Key

Bob Dennard's most profound contributions were to demonstrating the feasibility of MOSFET scaling, and then leading the way into implementation in real products. He worked with a succession of very talented engineers over several decades, providing guidance as well as continuing to make significant technical contributions.

The principles of scaling were first presented at the 1972 IEDM (4) along with the design and experimental characteristics of an ion-implanted 1μm transistor with a 20nm gate oxide[2] which had been optimized for scaling. One of the original slides used to describe scaling is shown in Fig. 1. (Bob remembers a high degree of skepticism about the feasibility of 20nm

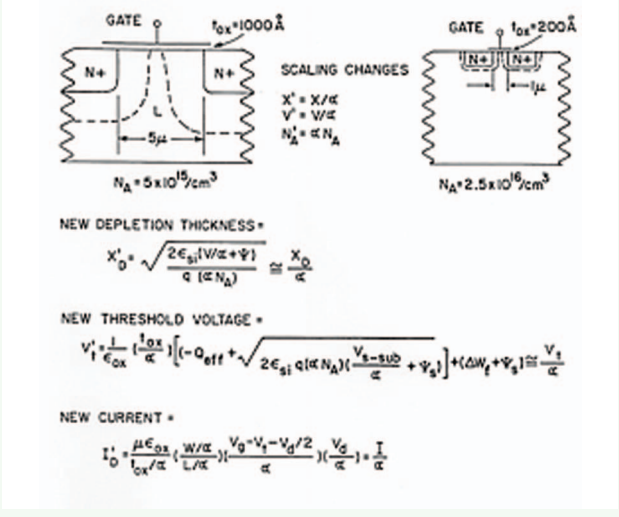

Fig. 1 Slide from 1972 IEDM showing some of the scaling principles.

oxides from the audience many of whom were struggling with making reliable 100nm oxides.) This was soon followed with a 1973 IEDM paper (6) utilizing ion-implantation to allow improved scaled transistors. The paper normally considered the "scaling paper" was published in 1974 (7). In 1975 Dennard, with others, proceeded to demonstrate scaling on a complex chip by scaling an existing 8Kb PMOS chip (originally designed in 3.75μm ground rules) by 3X and fabricating it with 1.25μm feature sizes using electron beam lithography (8). A photo of several cells and support circuits is shown in Fig. 2. Hwa Yu developed an anisotropic dry etching process which made it possible to delineate the 1.25μm features. The success of this experiment had a major impact on how seriously people took scaling both inside and outside IBM.

Attention was then turned to high-speed logic and SRAM. One of our goals was to lay the groundwork for replacing bipolar transistors in mainframe computers. This culminated in a series of eight papers (9) describing a 1μm technology that took advantage of the scaling principles. Bob was coauthor of several of the papers.

Bob continued to push the envelope with a large number of publications in cooperation with a succession of young researchers. Describing these papers is well beyond the scope of this paper. However, a few key papers stand out. In 1984, with Giorgio Baccarani and Matt Wordeman, he generalized the scaling theory to take into account the parameters which did not scale well (10). In 1985, he co-authored a definitive paper on 1μm CMOS (11) with Yuan Taur and others. In 1995, a paper laying the groundwork for a 0.1mm CMOS on SOI technology was published by Ghavam Shahidi and others (12).

In addition, Dennard furthered the cause and presented the challenges of MOSFET scaling to technical audiences outside the IEEE organization. For example, he published a paper in 1981 in the Journal of Vacuum Science and Technology (13) which showed the practicality of scaling to submicron devices and described the hierarchical wiring system needed to take advantage of scaling. In 1985 he published an authoritative paper on scaling to deep sub-micron dimensions in Physica (14).

Although he was not listed as an author, Bob had a major influence on the keystone 1988 paper (15) by Bijan Davari, et al, which described the 2.5V, 0.25μm CMOS technology which was key to the replacement of bipolar technologies for high-speed main-frame computers and microprocessors.

Technical Challenges and Advances to Make Scaling Feasible

Even though the principles of scaling, and the understanding that the MOSFET could be scaled existed in the early 1970s, the benefits of scaling could not have been accomplished without many other technical advances in the industry over the decades. There were remarkable improvements in optical lithography, dry etching, ion implantation, insulators, polycide and silicided contacts, multilevel metal, planarized BEOL, copper wiring, shallow trench isolation, packaging, design techniques, testing and characterization, design tools and system architecture. The switch to CMOS was critical to containing the level of chip power.

These improvements allowed scaling of the MOSFET technology to meet the expectations of the industry following the trends popularized in recent decades as Moore's Law (16).

The Long Delay before Switching to Lower Power Supply Voltages

While the advantages of scaling were apparent to many people, it was two decades before the power supply was scaled for mainstream products, Fig. 3. The indus-

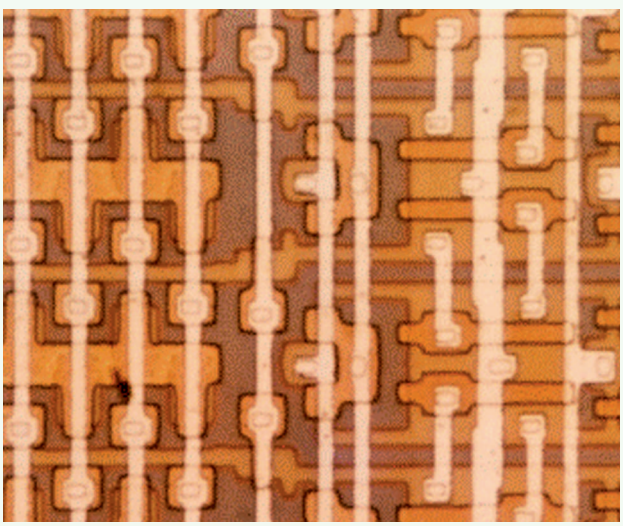

Fig. 2 Photograph of portion of experimental 8Kb DRAM chip using 1.25μm features which was scaled from a 3.75μm design.

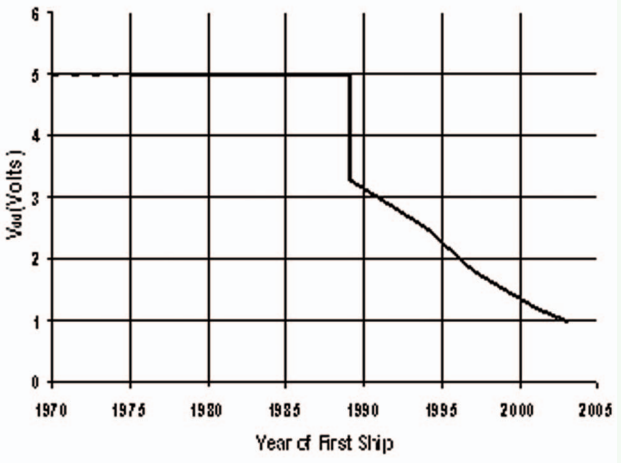

Fig. 3 Transition of mainstream MOSFET products from 5V to scaled voltages occurred two decades after scaling principles were defined.

try settled on 5V supplies in the early 1970s to be compatible with bipolar TTL. In fact, this was a lower voltage than what could have been possible for the dimensions being used. Consequently, the improvements in transistor design and chip fabrication were applied to 5V technologies, significantly improving component packing density and performance over several generations. Further, the LDD device (17) allowed reliable operation and high performance at 5V. The tighter tolerances necessary to make scaling practical improved 5V designs as well, reducing the performance advantage of full scaling. Most importantly, the whole computer industry was optimized around a 5V power supply and very successful products were being delivered. An earlier switch to a lower voltage would have been greatly disruptive to the designers, the manufacturers and in the marketplace.

The 5V standard finally collapsed in the late 1980s due to three major forces:

1) The power dissipation at 5V became untenable, especially as the circuits were driven to higher speeds.
2) The portable, battery-powered applications were demanding higher performance, low power and compatibility with battery voltages.
3) The inherent speed advantages of scaled transistors, as tolerances improved, were needed for high-speed applications.

Once the dam broke there was tremendous change within a few years, first to 3.3V then to 2.5V, etc.

The Impact of MOSFET Scaling has been Monumental

Scaled CMOS has become the dominant technology for digital and many analog applications and will continue to be a fundamental driving force of the industry for years to come.

By the late 1980s, DRAM had long displaced fixed head files in the file gap. In recent years, we have been seeing flash memory replacing disk drives in many portable applications.

The 2.5V CMOS technology (15) was the death knell for high performance silicon bipolar technologies in high-end computers. BiCMOS had gathered some momentum, but when designers came to realize that very effective off-chip drivers could be made using MOSFET circuits, BiCMOS soon faded. By the early 1990s, the high-end computers were being designed using low-voltage scaled CMOS (18) replacing bipolar chips. Bipolar and BiCMOS have found new applications for very high-speed applications using more exotic technologies.

Acknowledgements

The author is indebted to B. Davari, R. H. Dennard, E. J. Nowak and L. M. Terman for providing information for this paper. He also wishes to acknowledge a large number of outstanding engineers in IBM, other companies and Universities who shared an incredible 40 year journey in MOSFET technology.

References

[1] *To the Digital Age: Research Labs, Start-Up Companies, and the Rise of the MOS Technology*; Ross K Bassett, The John Hopkins University Press, 2002.

[2] "Design and characteristics of n-channel Insulated-gate Field-Effect Transistors"; D. L. Critchlow, R. H. Dennard, S. E. Schuster; IBM Journal of Research and Development, vol. 17, no. 5, p. 430, 1973.

[3] "An investigation of the potential of MOS transistor memories"; P. Pleshko and L. M. Terman; IEEE Transactions on Electronic Computers, vol. EC-15, No. 4, pp. 423-427, August 1966.

[4] "Design of micron MOS switching devices"; R. H. Dennard, F. H. Gaensslen, L. Kuhn, H. N. Yu; IEDM Tech. Dig., pp. 168 - 170, December 1972.

[5] "Fundamental limitations in microelectronics – 1. MOS technology"; B. Hoeneisen and C. Mead, Solid State Electronics, vol. 15, no. 7, pp. 819-829, July 1972.

[6] "Ion implanted MOSFETs with very short channel lengths"; R. H. Dennard, F. H. Gaensslen, H. N. Yu, V. L. Rideout, E. Bassous, A. LeBlanc; IEDM Tech. Dig., pp. 152 - 155, December 1973.

[7] "Design of ion-implanted MOSFET's with very small physical dimensions"; Robert H. Dennard, Fritz H. Gaensslen, Hwa-Nien Yu, V. Leo Rideout, Ernest Bassous, Andre R. LeBlanc; IEEE Journal of Solid-State Circuits, vol. 9, pp. 256 - 268, October 1974.

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

[9] "1 mm MOSFET VLSI technology: Parts I-VIII"; IEEE Journal of Solid-State Circuits, vol. SC-14, pp. 240-301, April 1979.

[10] "Generalized scaling theory and its application to a _ micrometer MOSFET design"; Giorgio Baccarani, Matthew R. Wordeman, Robert H. Dennard; IEEE Trans. Electron Devices, vol. 31, pp. 452 - 462, April 1984.

[11] "A self-aligned 1-μm-channel CMOS technology with retrograde n-Well and thin epitaxy"; Yuan Taur, Genda J. Hu, Robert H. Dennard, Lewis M. Terman, Chung-Yu Ting, Karen E. Petrillo; IEEE Journal of Solid-State Circuits, vol. 20, pp. 123 - 129, February 1985.

[12] "A room temperature 0.1μm CMOS on SOI"; G. G. Shahidi, C. A. Anderson, B. A. Chappell, T. I. Chappell, J. H. Comfort, B. Davari, R. H. Dennard, R. L. Franch, P. A. McFarland, J. S. Neely, T. H. Ning, M. R. Polcari, J. D. Warnock; IEEE Transactions on Electron Devices, vol. 41, issue 12, pp. 2405-2412, Dec. 1994.

[13] "CMOS scaling for high performance and low power – The Next Ten Years"; B. Davari, R. H. Dennard, and G. G. Shahidi; Proceedings of the IEEE, Vol. 83, No. 4, pp. 595-606, April, 1995. Earlier version published in Nikkei Microelectronics, pp. 144-1`54, September 1994.

[14] "MOSFET miniaturization – From one micron to the limits"; R. H. Dennard and M. R. Wordeman; Physica B + C, vol. 129, pp. 3-15, 1985.

[15] "A high performance 0.25 μm CMOS technology"; B. Davari, W. H. Chang, M. R. Wordeman, C. S. Oh, Y. Taur, K. E. Petrillo, D. Moy, J. J. Bucchignano, H. Y. Ng, M. G. Rosenfield, F. J. Hohn, M. D. Rodriguez; IEDM Tech. Dig., pp. 56 - 59, December 1988.

[16] Research Highlights with focus on Moore's Law; IEEE Solid-State Circuits Society Newsletter, vol. 20, no. 3, September 2006.

[17] "Fabrication of high-performance LDDFET's with oxide sidewall-spacer technology"; Paul J. Tsang, Seiki Ogura, William W. Walker, Joseph F. Shepard, Dale L. Critchlow; IEEE Trans. Electron Devices, vol. 29, pp. 590 - 596, April 1982.

[18] "Possibilities of CMOS Mainframe and its Impact on Technology R&D": A. Masaki, Symposium on VLSI Technology, May 28-30, 1991, pp. 1-4.

About the Author

Dale Critchlow is a retired electrical engineer with 35 years experience at IBM and 15 years in academia. He received his Ph.D. in Electrical Engineering from Carnegie Institute of Technology in 1956. After teaching at CIT for two years, he joined IBM Research. He became one the early members of the NMOS MOSFET project in the T. J. Watson Research Center in 1964, where he managed the MOSFET device and circuit design work through 1976. Next he transferred to the IBM Components Division, first in East Fishkill, NY, and then in Essex Junction, VT where he managed a group responsible for the advanced development of MOSFET logic and memory technologies. He retired from IBM in 1993 and was faculty member at the University of Vermont until 2005. He has active in IEEE activities and has published a number of papers and patents.

Dr. Critchlow is a Life Fellow of the Institute of Electrical and Electronic Engineers, an IBM Fellow and a member of the National Academy of Engineers.

[1] Ross Bassett wrote an excellent Ph.D. thesis and published a book [1] on the early history of the MOSFET technology. The appendices have a wealth of authoritative historical information.

[2] Concurrently, B. Hoeneisen and C. Mead published a theoretical paper [5] projecting that a 0.4mm transistor with 14nm oxides and 2V operation could be built.