Nanoscale CMOS

HON-SUM PHILIP WONG, SENIOR MEMBER, IEEE, DAVID J. FRANK, MEMBER, IEEE,
PAUL M. SOLOMON, FELLOW, IEEE, CLEMENT H. J. WANN, AND
JEFFREY J. WELSER, MEMBER, IEEE

Invited Paper

This paper examines the apparent limits, possible extensions, and applications of CMOS technology in the nanometer regime. Starting from device scaling theory and current industry projections, we analyze the achievable performance and possible limits of CMOS technology from the point of view of device physics, device technology, and power consumption. Various possible extensions to the basic logic and memory devices are reviewed, with emphasis on novel devices that are structurally distinct from conventional bulk CMOS logic and memory devices. Possible applications of nanoscale CMOS are examined, with a view to better defining the likely capabilities of future microelectronic systems. This analysis covers both data processing applications and nondata processing applications such as RF and imaging. Finally, we speculate on the future of CMOS for the coming 15–20 years.

Keywords*— *CMOS, device technology, memory, MOS, MOSFET, nanotechnology, scaling, ultralarge scale integration (ULSI), very large scale integration (VLSI).

I. INTRODUCTION

Silicon CMOS has emerged over the last 25 years as the predominant technology of the microelectronics industry. The concept of device scaling has been consistently applied over many technology generations, resulting in consistent improvement in both device density and performance. Device dimensions are now well below the micrometer scale and into the nanometer regime. The industry roadmap for CMOS technology development suggests that CMOS technology is nearing some fundamental physical limits in the not too distant future. It is therefore appropriate at this time to review the state-of-the-art, consider the potential capabilities of CMOS, and examine the fundamental physical limitations as well as practical technological barriers to continued development of CMOS technologies.

This paper will attempt to address the following questions. What are the apparent limits of CMOS technology? How can CMOS be extended into the nanometer regime? What sort of applications is it expected that CMOS will and will not be able to handle? The answers to these questions should create a background against which other nanotechnologies can be compared, perhaps making it clear where they can most usefully be applied. The paper begins by describing state-of-the-art CMOS technology, its expected evolution, and the apparent limits of scaling in Section II. Section III describes some novel possibilities for silicon technology at the limits of scaling, including both silicon-on-insulator (SOI) and its many variants, such as the ground plane FET and double-gate (DG) FET for logic, and single-element storage devices for memory. Section IV addresses the potential applications of nanoscale CMOS, including logic and memory, RF, and imaging applications. This discussion covers both the possibilities offered by nanoscale CMOS as well as the limitations imposed by such devices. This is followed by a discussion which summarizes the paper.

Manuscript received April 8, 1998; revised October 23, 1998.
The authors are with IBM T. J. Watson Research Center, Yorktown Heights, NY 10598 USA.
Publisher Item Identifier S 0018-9219(99)02365-8.

II. CMOS TECHNOLOGY OVERVIEW

To provide a background for discussing applications of nanoscale CMOS technology and potential nanoscale MOSFET innovations, we first briefly describe the "standard" model for continuing progress in CMOS and describe present state-of-the-art CMOS technology. These subjects are only covered briefly since they have been the subjects of many recent articles and reviews [1]–[3]. We also discuss some of the approaching difficulties along this path and describe recent results on low-power/low-voltage design and on fluctuations caused by the discreteness of dopants.

A. Scaling

The exponential growth in importance of CMOS in the last three decades has been largely driven by technological innovations that have enabled a steady reduction in MOSFET dimensions. The design of MOSFET's at progressively smaller dimensions has been by and large governed by the scaling criteria proposed by Dennard *et al.* [4] in the early 1970's. This scaling concept is illustrated in Fig. 1, which shows the larger device being scaled down by the factor α to yield the smaller device. According to simple electrostat-

0018–9219/99$10.00 © 1999 IEEE

TSMC Exhibit 1029

Table 1
Technology Scaling Rules for Three Cases (α Is the Dimensional Scaling Parameter, ϵ Is the Electric Field Scaling Parameter, and α_d and α_w Are Separate Dimensional Scaling Parameters for the Selective Scaling Case; α_d Is Applied to the Device Vertical Dimensions and Gate Length While α_w Applies to the Device Width and the Wiring)

Physical parameter	Constant-Electric Field Scaling Factor	Generalized Scaling Factor	Generalized Selective Scaling Factor
Channel length, Insulator thickness	$1/\alpha$	$1/\alpha$	$1/\alpha_d$
Wiring width, channel width	$1/\alpha$	$1/\alpha$	$1/\alpha_w$
Electric field in device	1	ϵ	ϵ
Voltage	$1/\alpha$	ϵ/α	ϵ/α_d
Doping	α	$\epsilon\alpha$	$\epsilon\alpha_d$
Area	$1/\alpha^2$	$1/\alpha^2$	$1/\alpha_w^2$
Capacitance	$1/\alpha$	$1/\alpha$	$1/\alpha_w$
Gate delay	$1/\alpha$	$1/\alpha$	$1/\alpha_d$
Power dissipation	$1/\alpha^2$	ϵ^2/α^2	$\epsilon^2/\alpha_w\alpha_d$
Power density	1	ϵ^2	$\epsilon^2\alpha_w/\alpha_d$

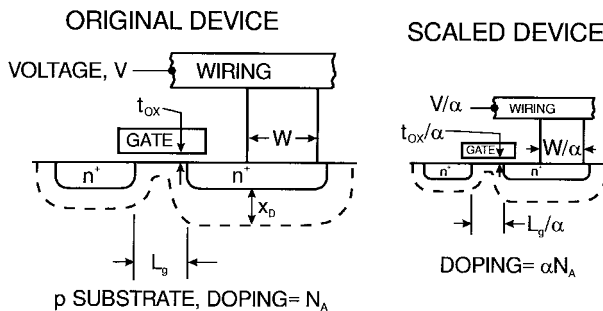


Fig. 1. Schematic illustration of the scaling of silicon technology by a factor alpha. Adapted from [5].

ics, if the dimensions, dopings, and voltages are scaled as shown, the electric field configuration in the scaled device will be exactly the same as it was in the larger device. These constant electric field scaling relations are summarized in column two of Table 1. There are two problems, however. The built-in potentials do not scale because they are tied to the silicon bandgap energy, which does not change (except by changing to a different semiconductor). Furthermore, the subthreshold slope cannot be scaled (except by lowering the temperature), since it is primarily determined by the thermodynamics of the Boltzmann distribution of carriers. Consequently, the threshold voltage cannot be scaled too far, or else leakage currents will become excessive. Both of these limitations cause deviations from simple scaling theory as the supply voltages approach 1 V.

In practice, because of these difficulties with low voltage, the voltage is not usually scaled as fast as the linear dimensions. This can be accommodated by introducing an additional scaling factor ϵ for the electric field (this ϵ is greater than one), as summarized under "generalized scaling" in column three of Table 1. Increasing the electric field requires increasing the amount of doping and also increases the power dissipation, but it does hold off some of the low-voltage difficulties. The disadvantage of this scaling is that the increasing electric field is a threat to the device reliability. Indeed, this reliability concern forces the use of lower supply voltages for smaller devices even when power dissipation is not an issue [1].

Finally, in recent generations of technology the wiring is not scaled to the same extent as the gate length, since this improves the wiring yield without degrading the gate delay. This approach, called "selective scaling," is shown in the final column of Table 1 and has two spatial dimension scaling parameters, α_d for scaling the gate length and device vertical dimensions and α_w for scaling the device width and the wiring. These approaches to scaling and issues related to them are described in more detail in [1].

The expected power densities and delays of future technology generations have been estimated using these selective scaling rules and are illustrated in Fig. 2 down to near the limits of scaling [1], [5]. The high performance option yields high logic speeds, with loaded delays down to 80 ps, even for static CMOS. Still higher speeds are expected for dynamic logic families, but as shown in Fig. 2(a), these high speeds are at the expense of high power densities. Note that even though the lower voltages of the low power path result in an initial savings in power, the power density starts rising on that path, too, for gate lengths below 0.25 μm.

B. Industry Projections

Scaling theory in conjunction with observations of past industry trends (e.g., "Moore's Law") has led to the creation of so-called "roadmaps" for CMOS technology, the most

TSMC Exhibit 1029

Table 2
Highlights of the 1997 Semiconductor Industry Association National Technology Roadmap for Semiconductors [6]

Year first ship	1995	1997	1999	2001	2003	2006	2009	2012
DRAM (bits/chip)	64M	256M	1G		4G	16G	64G	256G
DRAM chip size (mm^2)	190	280	400	445	560	790	1120	1580
μP transistors/cm^2		3.7M	6.2M	10M	18M	39M	84M	180M
μP chip size (mm^2)	250	300	340	385	430	520	620	750
Gen. Lithography (μm)	0.35	0.25	0.18	0.15	0.13	0.10	0.07	0.05
Gate Lithography (μm)	0.28	0.20	0.14	0.12	0.10	0.07	0.05	0.035
Oxide thickness (nm)	7-12	4-5	3-4	2.4-3.2	2-3	1.5-2	<1.5	<1.0
Supply voltage (V)	3.3	1.8-2.5	1.5-1.8	1.2-1.5	1.2-1.5	0.9-1.2	0.6-0.9	0.5-0.6
V_T 3σ variation ($\pm mV$)	60	60	50	45	40	40	40	40
Clock (MHz) (across chip)	300	750	1200	1400	1600	2000	2500	3000

public of which is the Semiconductor Industry Association Roadmap, some highlights of which are shown in Table 2 [6]. Note that the effective channel length is expected to be 10–30% less than the gate lithography, depending on the manufacturer.

The numbers on this roadmap are based on simple projections of past progress. These are simply targets that the industry intends to try to meet. In many cases it is not known how the targets can be met and in some cases there are not even any very good ideas. For example, for DRAM, these numbers require gradually reducing the area required for a single bit down to four lithographic squares, which will require serious innovation, since there is no obvious present path to accomplish this. Several potential exploratory approaches to this problem are discussed in Section III-B.

The anticipated increase in chip sizes will both require and accompany increases in wafer sizes. 300-mm diameter silicon wafers are expected to be used in production starting around the turn of the century, and still larger wafers (perhaps 450 mm) are being considered for the future.

The economic importance of CMOS has been scaling up exponentially at the same time device dimensions have been scaling down exponentially. Fig. 3 shows historical data for revenues for the largest CMOS markets (DRAM's and microprocessors) and their extrapolation into the future. It also shows the trend lines for progressively larger industry groupings. As can be seen, both components of the CMOS market have been growing at an average annual compound growth rate (CGR) of about 25%, which is much faster than the industry groups of which they are a part. If the CMOS industry continues to expand at its present rate, the entire semiconductor industry CGR will have to increase in the next decade as CMOS grows to completely dominate it, and in the following decade the entire electronics industry would come to be dominated by CMOS. By the mid-2010's, annual CMOS revenues in the trillions of dollars would account for roughly 10% of the gross world product (GWP). Continued extrapolation would be even faster than that considered by Moore [7], reaching the entire GWP by the late 2020's. Such rapid change of the entire world economy is almost certainly impossible, implying that the industry growth rate will have to slow during the next 10–20 years to a rate commensurate with the GWP.

One final economic concern for the future of CMOS is the rising cost of fabrication facilities [7]. The cost of new fabs has been rising quite quickly, but this is partly because their size has also been increasing to accommodate the increasing demand for silicon. When normalized to depreciation dollars/cm^2, recent data on fab cost, though fairly widely scattered, shows an annual CGR of 8–10%, which appears consistent with long-term historical data. On average, this has been paid for by increasing chip prices (dollars/cm^2) of 5–6% per year coupled with continuing incremental improvements in manufacturing efficiency. These improvements in efficiency have presumably enabled the slow increase (3%/year CGR) in the fraction of revenues reinvested in capital improvements by the semiconductor industry. On the basis of these considerations, it does not appear likely that the cost of fabrication facilities will be a major determining factor for future CMOS technology, unless future generations of technology deviate significantly from past trends. Significant efficiency improvements cannot continue indefinitely, but at their historically low rate they may be able to continue for quite some time, and even if they do not, the impact on the overall dynamics of the industry would probably be tolerable.

C. Device Structure and Technology

1) Historical Development: It is rather amazing that CMOS has had the same basic device structure through decades of development. The seemingly straightforward

TSMC Exhibit 1029

Table 3
Milestones in Bulk CMOS Technology [This Table Reports the Technological Advances Which Have Made It into Today's CMOS Production Process, Usually By Their Date Of First Appearance at the IEDM, in the Context of MOS Very Large Scale Integration (VLSI)]

Year	Technology	Channel Length (μm)	reduction to practice	Reference
1960	MOSFET	-	lab	[172]
1965	64b MOSFET SRAM	-	lab	[172]
1966	polysilicon gate self-aligned FET	-	lab	[173]
1968	one transistor memory cell	-	lab	[32]
1969	ion-implanted channel	-	lab	[173]
1970	CMOS Watch chip	-	product	[172]
1971	Intel 4064 Microprocessor	10	product	[174]
1979	silicided polysilicon gate	1	IEDM	[175]
1980	sidewall spacer for S/D implant	1.5	IEDM	[14]
1982	Self-aligned silicide (salicide)	-	IEDM	[176]
1982	Trench Isolation	-	IEDM	[177]
1983	oxy-nitrides for gate dielectric	-	lab	[178]
1985	halo doping of source/drain	0.2	lab	[179]
1986	N^+ - P^+ polysilicon gates	0.5	IEDM	[180, 181]
1986	retrograde channel doping	0.5	IEDM	[180]
1987	0.1mum MOSFET	0.1	IEDM	[11]
1989	chemical-mechanical polishing	-	IEDM	[182]
1992	Damascene Interconnect Technology	-	IEDM	[183]
1993	Copper Interconnect	-	IEDM	[184]

path of CMOS scaled toward 100 nm has in fact involved a tremendous amount of technological innovation not obvious in the device structure. Table 3 summarizes major milestones of structural and technological advances of CMOS.

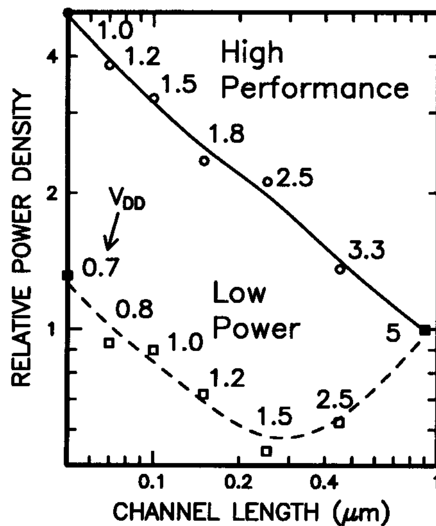
The basic self-aligned polysilicon gate MOSFET was introduced in 1970, with ion implantation being used subsequently in the 1970's for the source and drain regions, and soon after for the channel as well. By the end of the 1980's, silicided polysilicon gate self-aligned CMOS was fast becoming the industry standard. As discussed before, the scaling criteria for MOSFET's was proposed in the early 1970's [4] and by and large this has been adhered to, except for voltage which has not been reduced as fast as stipulated (proportional to gate length) in the interests of voltage standardization and performance. Hot carrier reliability issues were the main driving force behind reduced power supply voltages beyond the sub-0.5 μm generation.

One major difference which sets the submicron MOSFET apart from its predecessor was the introduction of the sidewall spacer, enabled by reactive ion etching. Other changes include self-aligned silicide, which was introduced in the mid-1980's; shallow trench isolation, which is replacing all the variations of local oxidation of silicon (LOCOS) [8], and chemical-mechanical polishing (CMP)-based processes, which took over the metallization process. There were also more subtle evolutions: substrate engineering became widely popular in the 0.25-μm regime [9], [10], as did source drain and halo engineering in 0.1-μm generations [11]–[13], though the idea was introduced much earlier in the form of lightly doped drain (LDD) to reduce hot carrier-induced device degradation [14]. Even the role of the spacer also evolved with time: the spacer was first introduced to realize LDD. After the industry decided it could freely scale the internal power supply voltage, LDD was abandoned, but the spacer stayed and became a must for salicide processes.

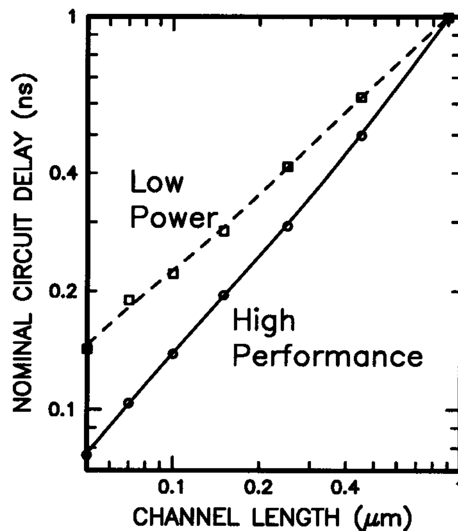
The 100-nm barrier was broken by Sai-Halasz *et al.* [11], [15] in 1987. By 1997, there were many research articles describing sub-100-nm gate length MOSFET's [12], [13], [16]–[24].

2) State of the Art: Several recent review articles have summarized the characteristics of state-of-the-art CMOS technology [1]–[3]. Fig. 4 illustrates most of the impor-

TSMC Exhibit 1029



(a)



(b)

Fig. 2. (a) Relative power density versus effective channel length for high performance and low power technology scaling and (b) corresponding loaded circuit (three-input NAND) delay versus channel length for the same technology scalings. Values are mostly from [1]. Adapted from [5].

tant features. The gates are fabricated with n- or p-type polysilicon so that both nFET's and pFET's are surface channel devices, for maximum performance. The gates are topped with a metal silicide for lower gate series resistance, although the resistance is still higher than would be desirable for maximum RF performance (see Section IV-B1). Special lithographic techniques are used to pattern the gates with minimum dimensions 20–30% below the general lithographic feature size. The gate dielectric must be very thin, typically around 3 nm for the 0.1-μm gate length generation of technology. Scaling requires such thin oxides to adequately limit short channel effects, and to provide sufficient current drive.

Shallow trench isolation between devices involves etching trenches, filling with deposited oxide, and polishing to planarize. This process allows devices to be placed much closer together than was formerly possible with local oxidation of silicon (LOCOS), resulting in higher circuit density. The source and drain use shallow, moderately doped extensions under the gate edges and gate sidewalls. These are engineered to reduce short channel effects and series resistance and yet provide adequate reliability with regard to hot electrons, while the deeper contact implants allow low-resistance silicide contacts.

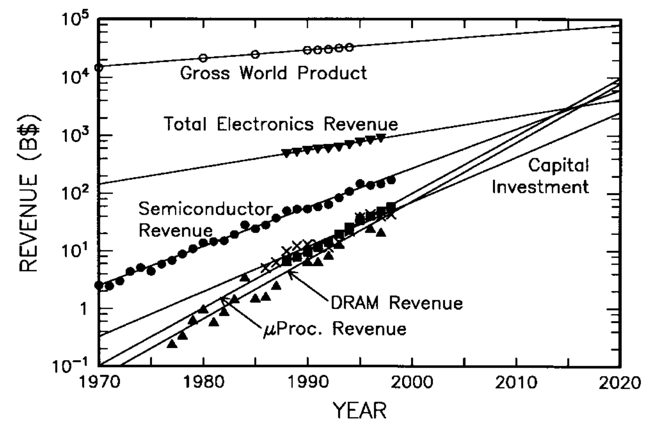


Fig. 3. Plot of revenue versus year, with projections into the future, for the CMOS industries, the semiconductor industry, the electronics industry, and the entire world economy. Data are from Dataquest and ICE. GWP data are courtesy of M. Moser.

Of great importance in achieving the shortest possible channel lengths is the engineering of the doping profiles in the channel region. Retrograde doping profiles can reduce transverse electric field in the channel (improving mobility), while at the same time reducing two-dimensional (2-D) effects by shielding the drain potential from the channel region. Halo implants [2] at the source and drain edges can be used to at least partially cancel 2-D-induced threshold voltage shifts (V_T), resulting in less V_T roll off.

The drawing in Fig. 4 does not show the wiring levels, but the wires are obviously very important in creating large integrated circuits (IC's), and substantial technological progress is occurring there, too [25], [26]. By the year 2000, it is anticipated that most of the wiring will use copper for its lower resistivity (40% lower than Al/Cu), lower processing cost, and reduced electromigration [27], [28]. Lower dielectric constant insulators are also being considered to reduce wiring capacitance and improve speed. Materials such as spin-on-glass (FOX, Xerogel) and some polymers have dielectric constants ranging from 3 to 1.8 [26], [29], [30] and are actively being investigated for use as intermetal dielectrics. The use of a hierarchy of wiring sizes, from very fine wires at minimum lithographic dimension on the bottom to large "fat" wires on the top, is expected to prevent wiring from becoming a major bottleneck in future IC's [31].

While the focus for logic device development has been mostly on increasing performance, memory improvements have largely been measured by increases in density. In fact, the strong economic pressures to decrease the cost per storage bit, caused by the size and commodity nature of the DRAM market, has been the major motivator for the rapidly

TSMC Exhibit 1029

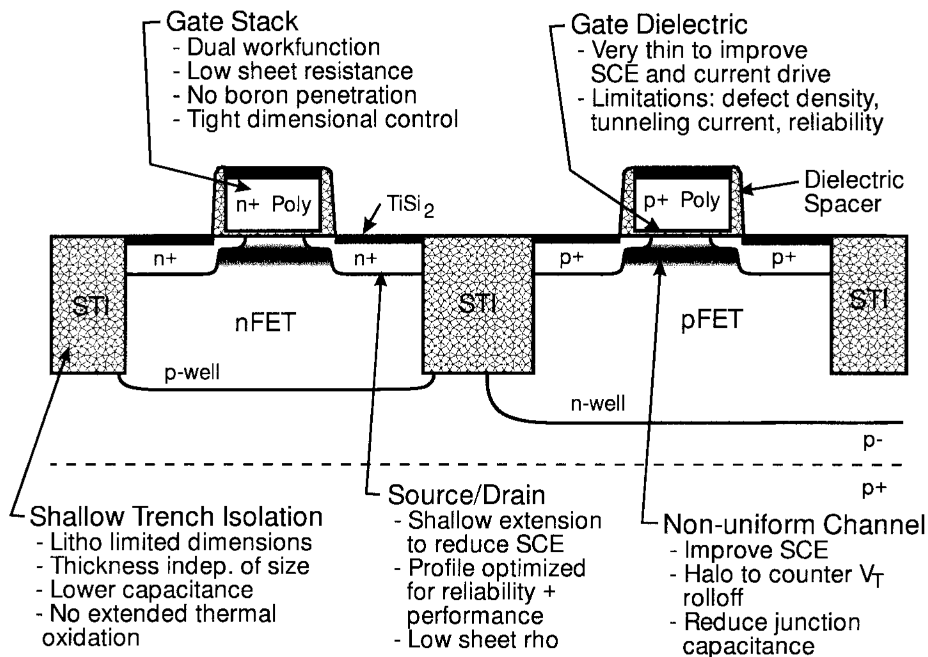


Fig. 4. A cross-sectional diagram of typical state-of-the-art CMOS technology, indicating some of the more important features. Adapted from [1] and [5].

shrinking lithography feature sizes, which has also benefited logic. The basic one-transistor/one-capacitor cell [32] has remained the industry standard since the early 1970's, but increasingly complicated structures and processing have become necessary to keep up with the historical roadmap, which has produced a new generation with a four times bits/chip increase every three years. The current generation of widely available DRAM have 16 or 64 Mb per chip, with 256 Mb chips soon to follow.

D. Challenges for CMOS Technology

There are many technological challenges facing CMOS in the near future, some of which have been alluded to already. The SIA roadmap [6], for example, labels many of its projected industry requirements as having no known solution. In this section, we highlight some of the key device and circuit challenges facing nanoscale CMOS.

1) Device Technology: Referring to Fig. 1, the critical dimensions that need to be engineered are the gate length (L_g), the gate oxide thickness (t_{ox}), the depletion depths under the gate (x_{depl}), the source/drain junction depth (x_j), and steepness of the source/drain doping profile. All these quantities must be scaled together.

a) Gate length: The gate length is the smallest feature of the MOSFET patterned by lithography and etching. Optical lithography has been able to provide generations of feature size reduction mainly through reduction of the wavelength of light employed [33]. Employing light with wavelength shorter than 193 nm (for gate lithography in the 2003 generation) presents many difficulties, among which the availability of materials for the optical system is a major barrier. Currently there is much debate about how the lithographic requirements will be met beyond the 2001 time frame, with X-ray, extreme UV, and e-beam all being considered. Sublithographic feature size may be obtained by etching techniques or sidewall image-transfer techniques. However, these largely experimental techniques have never been proven in a manufacturing environment, nor do they seem to improve line pitch. The development of a reliable, manufacturable, cost-effective lithographic technique is absolutely essential to continued progress in CMOS technology.

b) Gate oxide thickness: As indicated in Table 2, the electrical thickness of the gate insulator must decrease with the channel length. Recent studies of tunneling through thin oxides [34]–[36] have shown that silicon dioxide can potentially be thinned down to slightly below 2 nm before the leakage current and associated dissipation become so large as to be unacceptable (Fig 5). Note that tunneling occurs not only to the inversion layer, but also to the accumulation layers, which exist where the gate overlaps the source and drain regions. As device areas are scaled, this latter component can be appreciable [37], particularly since it flows whenever the drain of the device is biased, even if the gate is off. For equivalent (electrical) SiO_2 thicknesses below 2 nm, thicker gate insulators with a higher dielectric constant than silicon dioxide are being considered as a way to reduce the tunneling current through the gate insulator, though the thickness of the insulator cannot grow too much without interfering with the scalability of the FET due to 2-D effects in the physically thicker gate insulator [38]. In addition, reliability and insulator/semiconductor interface properties remain the most important concerns for such new materials.

c) Depletion depths and junction depths: Depletion depths and junction depths have traditionally been controlled by ion-implantation of appropriate dopants into selected regions and limiting their movement during

TSMC Exhibit 1029

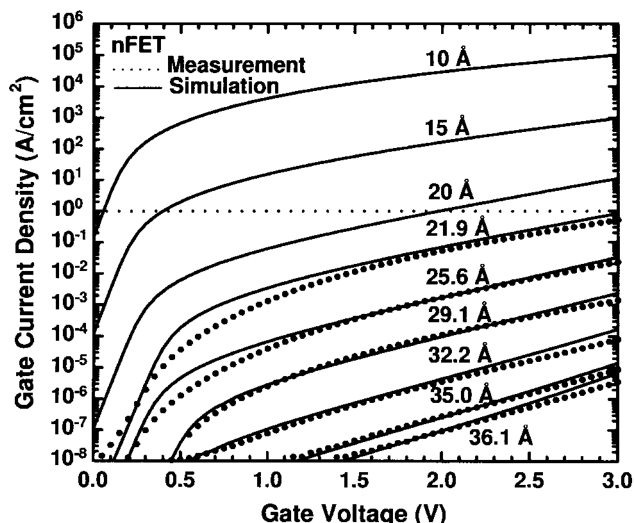


Fig. 5. Calculated (lines) and experimental (dots) results for tunneling currents from inversion layers through thin oxides. Adapted from Lo *et al.* [35].

subsequent heat cycles. Future generations of technology require ever steeper doping profiles. The sub-50-nm regime will apparently require profiles of order <5 nm/decade. Present annealing procedures are unable to produce such steep profiles, but a method must be found if such devices are ever to be manufactured.

Another related concern is that the very thin depletion depths needed in future CMOS require very high body doping concentration, perhaps into the 5×10^{18} cm^{-3} range for sub-50-nm FET's. At these doping levels direct body-to-drain tunneling leakage is expected to become a serious problem.

By changing the device structure to use a very thin undoped silicon channel such as the double-gate (DG) MOSFET [39], [40] (see Section III-A), precise dopant placement in the channel region is no longer necessary. However, the need for a steep lateral junction dopant profile remains. The thin undoped silicon channel also alleviates the discrete dopant fluctuation problem (see Section II-D3). However, confining the channel to a thin silicon layer introduces two potential problems: 1) quantization modulated threshold voltage and 2) source/drain resistance due to depletion of the source/drain. The former is discussed further in Section III-A6. The "heavily doped" drain can easily be depleted by the gate to drain bias. For an equivalent gate oxide of 1 nm, $V_G - V_T = -0.4$ V, $V_D = 1$ V, the depleted silicon depth is $\simeq$3 nm for a drain doped to 10^{20} cm^{-3}. A DG MOSFET therefore must have a channel at least 6-nm thick or a source/drain doping greater than 1×10^{20} cm^{-3} to avoid current degradation due to drain depletion.

2) Memory Technology: Though this paper will not focus on DRAM technology [41]–[43], it is worth mentioning a couple of the difficulties being faced as the gigabit generation is reached. One major challenge of all generations has been to maintain adequate capacitance for charge storage with decreasing capacitor footprints. To this end, three-dimensional (3-D) capacitors, consisting of either deep trenches dug into the silicon substrate or conductor/dielectric stack structures fabricated above the transfer transistor, have been used since the 4-Mb generation [44]. However, at gigabit densities it is becoming increasingly difficult to maintain the capacitance of these structures. In addition, the folded bit-line memory array architecture creates a theoretical lower limit on the cell size of 8 F^2 (where F is the minimum lithographic feature size in a given technology) due to wiring of two word-lines and one bit-line per cell. Moving to an open bit-line architecture, with a wiring limit of 4 F^2, will only increase the pressure to maintain high capacitance as noise levels increase. Furthermore, the performance of DRAM chips continues to lag further behind the processors they are meant to serve, increasing the desire toward higher performance DRAM, and the integration of memory and logic on the same chip to increase bandwidth [45].

The industry will no doubt continue to push DRAM capacity through several more generations, but ultimately a different memory concept will be necessary for true nanoscale storage. A possible candidate for this is the floating-gate transistor, which is currently used in non-volatile FLASH/EEPROM chips [41], [46]. This memory cell consists of only a single transistor, so it is highly scalable, and since it does not require complicated capacitor fabrication, it is highly integratable with logic. The prospects for nanoscale FLASH devices will be discussed in Section III-B.

3) Random Fluctuation of Device Properties: Random fluctuation of device properties may ultimately limit the number of devices which can be integrated on one chip. Fluctuations of device properties result in variations of transistor current drive capabilities and propagation delays, leading to intolerable clock skews or malfunction of circuits that depend on matched or absolute values of device properties. The random fluctuation of the threshold voltage of a MOSFET due to the random fluctuation of the placement and number of dopant atoms is perhaps the most studied example associated with fundamental physics [47]–[51] (as contrasted with fluctuations arising from the finite ability to control the fabrication process) and will be examined in this section.

Keyes [48] provided the first theoretical model of threshold voltage fluctuation for a uniformly doped MOSFET using percolation theory. Burnett and Sun [52] studied experimentally the bit-fail rate of SRAM cells (where a matched pair of minimum sized transistors are used) as the operating voltage was reduced by examining the threshold voltage of the failed bit cell and found general agreement between Keyes' theory [48] and experiment. In a more simplified model, to first order, accounting only for the fluctuation of the number of dopant atoms and ignoring the placement of the dopant atoms [49], the standard deviation of the threshold voltage can be described by [2], [53] $\sigma_{V_T} = (q/C_{\rm ox})\sqrt{(N_A W_{dm})/3LW}(1-(x_s/W_{dm}))^{3/2}$, where $C_{\rm ox}$ is the gate oxide capacitance, L and W are the length and width of the MOSFET, W_{dm} is the maximum gate-induced depletion depth, x_s is the depth of the low-impurity

TSMC Exhibit 1029

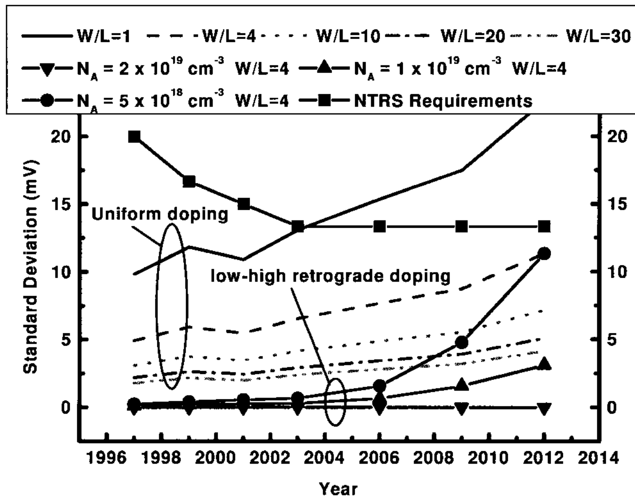


Fig. 6. Standard deviation of the threshold voltage of MOSFET's scaled according to the SIA roadmap. The solid line with square symbols shows the standard deviation (one sigma) of the threshold voltage as prescribed by the SIA roadmap. The lines (without symbols) show the standard deviation for the uniform doping case for various transistor W/L ratios. The thin solid lines with solid circles and triangle symbols are the standard deviation (one sigma) for retrograde channel (low–high) doping with the high doping (N_A) at 5×10^{18} cm^{-3} (circle), 1×10^{19} cm^{-3} (up triangle), 2×10^{19} cm^{-3} (down triangle), respectively, for $W/L = 4$. The maximum depletion depth (W_{dm}) is kept the same in these cases. Adapted from Wong *et al.* [53].

(assumed undoped) region at the surface, and N_A is the doping concentration of the heavier doped channel below the low-doped region. For a MOSFET with $W/L = 1$ and $L = 100$ nm and a uniform doping, the standard deviation of the threshold voltage is 15 mV. This result agrees qualitatively with the more rigorous 3-D simulations of MOSFET with a random dopant distribution [49].

Using the simple model above, the standard deviation of the threshold voltage of MOSFET's scaled according to the SIA roadmap [6] (see Table 2) can be estimated and is shown in Fig. 6. The lines show the results for the uniform doping case for various transistor W/L ratios, where the doping increases with each generation to achieve the desired V_T. While the uniform doping case is analytically simple to analyze, the more interesting cases are for low–high (retrograde) doping where the effects of dopant fluctuations are reduced because dopants are removed from the channel region. The thin solid lines with solid symbols show the standard deviations (for $W/L = 4$) obtained for a low–high doping with various high doping levels (N_A). The maximum depletion depth (W_{dm}) is fixed within each generation but varies between generations to satisfy V_T requirements. These results show that low-high doping is effective in suppressing dopant fluctuation effects if a "ground-plane" like sharp doping profile can be achieved with the ground-plane doping (N_A) greater than 1×19^{19} cm^{-3}.

The solid line with square symbols (Fig. 6) shows the standard deviation (one sigma) of the threshold voltage as prescribed by the SIA roadmap. It appears that without the "ground-plane" like doping, narrow devices will be unable to meet the roadmap requirements, while even for wider FET's the threshold voltage uncertainty due to dopant number fluctuation alone will account for a significant fraction of the V_T tolerance budget beyond the 70-nm generation, which places an unrealistic expectation on the ability to control process-related parameter fluctuations.

Fig. 7 shows the maximum statistically expected intra-chip deviation of the threshold voltage, which is computed by $\Delta V_T(\max) = \sqrt{2}\mathrm{erf}^{-1}(1 - (1/N_{\mathrm{transistors}})) \times \sigma_{V_T}$, where $N_{\mathrm{transistors}}$ is the total number of transistors on the chip. The maximum expected deviation indicates how far off the threshold voltage can be in a large chip, but it does not provide a simple correlation with circuit functionality. Information such as the number of critical paths and the sensitivity of critical performance gauges to threshold voltage variations must also be considered when evaluating circuit functional yield. For digital circuits, it is fair to say that as the power supply voltage becomes close to the threshold voltage, the effect of threshold voltage fluctuation would be more severe. This is illustrated further in Fig. 7 by plotting an estimated maximum allowable V_T deviation, which is taken to be a percentage of the gate overdrive, $V_{DD} - V_T$, for each generation. The exact value of this percentage (8.3%) was chosen to make the maximum allowable deviation equal to the maximum expected deviation for the 1999 technology node. The maximum allowable deviation computed from $V_{DD} - V_T$ decreases more rapidly than the maximum expected deviation targeted by the SIA roadmap, creating an inconsistency. If circuit functionality is tied strongly to $V_{DD} - V_T$ as assumed, Fig. 7 suggests that either the threshold voltage tolerance target should be further tightened or $V_{DD} - V_T$ should be maintained at a level higher than projected.

It should be remembered that V_T fluctuation due to discrete dopants is only a portion of the total V_T variation, since other process-induced fluctuations of the threshold voltage may well be substantial. Also, the threshold voltage fluctuation due to dopants is not completely determined by the number of dopants in the depletion region under the entire gate of a uniformly doped channel. 3-D random dopant simulations [49] show that device characteristics are asymmetric upon reversing the source and drain terminals, indicating that the placement of dopants near the source/drain regions plays a significant role in determining device characteristics, especially at high drain voltages. For deep submicron MOSFET's, channel potential control is typically aided by halos or pocket implants locally around the source/drain extensions [2]. Local dopant fluctuation effects have yet to be quantified in ongoing research, most probably through 3-D device simulations similar to the work in [49].

In order to avoid uncertainties due to random dopant fluctuations, nanoscale MOSFET's are often designed with a thin, undoped silicon channel. Threshold voltage control is then dependent on the workfunction of the gate material. This is described in more detail in Section III-A6.

4) Low Power Considerations: The earlier discussion of scaling pointed out the rising power density associated with

TSMC Exhibit 1029

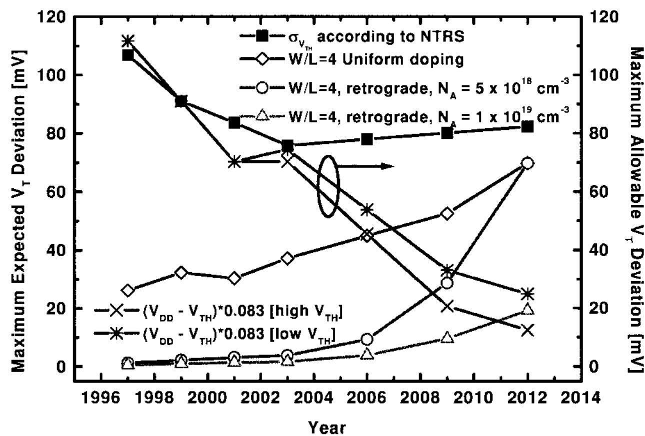


Fig. 7. Maximum statistically expected intrachip V_T deviation for MOSFET's scaled according to the SIA roadmap. The open symbols are for discrete dopant fluctuations only, using $W/L = 4$, for uniform doping (diamond) and low–high retrograde doping of $N_A = 5 \times 10^{18}$ cm^{-3} (circle), and $N_A = 1 \times 10^{19}$ cm^{-3} (up triangle). The square symbols are the maximum expected V_T deviation based on the σ_{V_T} targeted by the SIA roadmap, which includes fluctuation due to all sources, not just discrete dopants. The number of transistors in a chip is assumed to be the number of transistors in a microprocessor chip. The right axis shows the estimated maximum allowable V_T deviation for each technology generation. This maximum allowable deviation is taken to be a percentage of the gate overdrive, $V_{DD} - V_T$, for each generation. The exact value of this percentage (8.3%) was chosen to match the maximum allowable deviation requirement to the maximum expected deviation for the 1999 technology node. Adapted from Wong *et al.* [53].

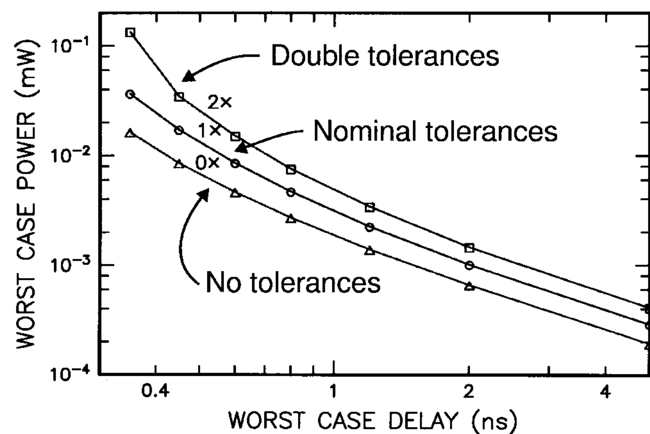


Fig. 8. Plot of optimized power versus delay for three different tolerance conditions. The one times conditions are as follows: global (chip-to-chip) Gaussian variations with $1\sigma = 6\%$, 2% and 10 mV for channel length, V_{DD} and V_T, respectively, and local (device-to-device) Gaussian variations with $1\sigma = 5\%$, 4%, and 7 mV for channel length, V_{DD} and V_T, respectively. The channel length variation also causes additional V_T variation due to the short channel effect model being used (see [61]).

future generations of technology. If past trends continue and designers keep making larger and larger chips with each successive generation of technology, this power dissipation threatens to inhibit seriously many possible applications, especially portable applications intended to run on batteries. Low-power design techniques are intended to ameliorate this problem and have been the subject of many recent articles (see, for instance, [54]–[57]). The most effective place to reduce power dissipation is almost always at the highest level of the problem definition. Redefining the problem, the architecture, the algorithms, and/or the protocols can often save several orders of magnitude in power dissipation. The development by Meng *et al.* [58] of a portable video-on-demand chip set using only 10 mW is an example of this.

At the technology level, the most effective way to reduce the power is to lower the supply voltage, as was illustrated in the low-power design points in Fig. 2. Various workers have considered the question of how to choose the optimum threshold voltages for low-power design, both with [59]–[61] and without [62]–[64], taking into account device and process variations. Optimizations without these variations can be misleading, however, since the variations play an increasingly important role in circuit performance at low voltage, as described in Section II-D3 and demonstrated experimentally in [65] for ring oscillators. The recent optimizations of Frank *et al.* [61] are perhaps the most detailed in this area, taking into account both device-to-device and chip-to-chip parameter and supply variations, including short channel effects, and using full circuit simulations to obtain accurate speed and power information.

Typical results for these optimizations are shown in Fig. 8. These particular optimizations are for static CMOS arithmetic circuits, and each point in the figure represents an independent optimization of both the supply voltage and the threshold voltage. Optimization of the gate length was also considered but was found to be a weaker effect. These power-delay curves are for 0.1-μm channel length and show power varying as $\tau^{-1.3}$ to τ^{-3} as the delay varies from long to short, indicating that the best low power tradeoffs occur for fast circuits. These curves cover a much wider range of power-delay tradeoffs than the two cases shown in Fig. 2, both of which are at the fast end of the scale. Note that even in the slow regime, the energy per operation (proportional to the product of power and delay) is still decreasing, albeit slowly.

These studies show that even in the presence of realistic process and supply variations the optimum supply voltages can readily drop below 1 V and can reach 0.5 V if the speed target is slow enough. The optimum value for V_T increases for slow circuits, to reduce static dissipation, and increases by 20–100 mV when the tolerances are doubled from their nominal values to the two times values used for the highest power case in Fig. 8, with the largest increases occurring for the shortest gate lengths. The dependence of the optimum design points on activity factor and logic depth is illustrated in Fig. 9, where the parameter tolerances correspond to the one times case in Fig. 8. As shown, the optimum nominal threshold voltage can become quite low for high duty factor and/or short logic depth circuits. These conditions are not expected to be common in low-power circuitry but may be important in high-performance designs.

5) Perceived Limits of CMOS Technology: The ultimate limit of CMOS technologies has been the subject of

TSMC Exhibit 1029

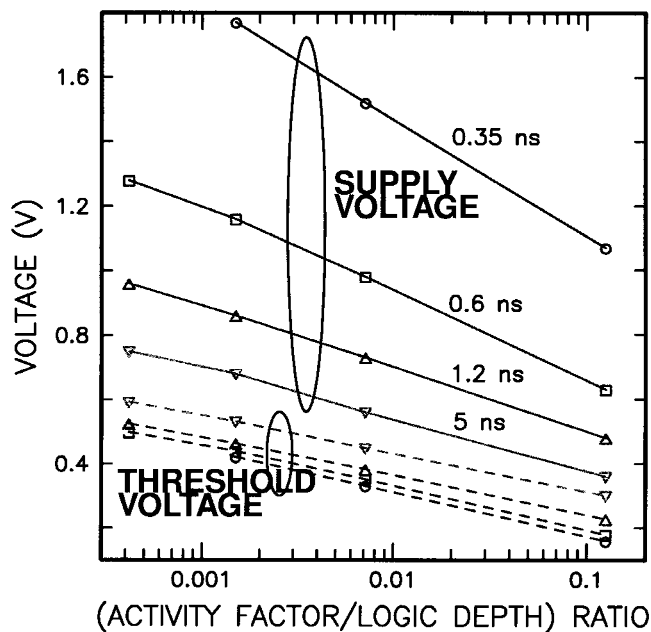


Fig. 9. Plot of supply voltage and threshold voltage versus the activity factor-to-logic depth ratio for four different delay constraints. Logic depth $= 4n_{\mathrm{stages}} = 48$, 40, 28, and 8 for the data points from left to right across the plots. Data are from [61].

discussion since the early 1970's [47]. Although all the technological predictions in the past have proven to be too pessimistic, there is a feeling, as we approach the turn of the century, that CMOS will be facing some fundamental physical limits in the not too distant future. Meindl *et al.* [66]–[68] aptly summarized these limits as a hierarchy of limits: systems; circuit; device; material; and fundamental physics. Many papers were published on this subject in recent years [1]–[3], [25], [66]–[70].

From a device design point of view, the depletion depth for bulk CMOS is limited to about 10 nm due to body-to-drain tunneling current limitations on maximum body doping and to difficulty in controlling the extremely abrupt doping profiles necessary. For the more advanced structures described in Section III-A, quantum mechanical effects would render control of threshold voltage difficult below 5 nm channel thickness. From the work of Taur *et al.* [2], the gate oxide thickness (for SiO_2) would be limited to about 2 nm before direct tunneling current (hence power dissipation) become excessive even for high-performance logic. Assuming that the lithography and patterning capabilities will deliver the required linewidths, the minimum channel length would be limited to about 50 nm for bulk CMOS and to about 25 nm for DG CMOS [2], [39].

III. Silicon Technology at the Limits of Scaling

A. Logic Devices

While CMOS is on its fast scaling track, alternative technologies are hard put to compete since every three years sees a two times increase in performance and a four times increase in the number of devices per chip (Moore's law). With the perceived end of conventional CMOS scaling on the horizon, advantages inherent in other, albeit more expensive, technologies and design approaches may have a greater impact. Perhaps some other approach, being experimented with today, may well become the main line logic and memory technology of the future.

For the remainder of this discussion, we will focus on SOI and its variants since we believe that this is the paradigm that will take over once the progress of bulk CMOS slows down. The reason for this is that SOI uses the same substrate, the same material set, and the same fabrication processes as bulk CMOS, yet it embodies an extra degree of freedom, almost perfect isolation, which can be used to gain higher performance, lower power, and other desirable attributes.

SOI, being still in the preproduct stage, has many variants, and one can discern an evolutionary path, within SOI, from very bulk-like, to very venturesome structures. This path is illustrated in Table 4.

1) SOI Wafers: The SOI starting material is usually made by one of the two processes: SIMOX or bonded SOI (BESOI) [71]. For SIMOX, oxygen is implanted at sufficiently high density that it can react with silicon, during a post-implant high temperature bake, to form a continuous silicon dioxide film buried under a thin silicon surface layer. This is now the leading commercial process for making SOI. The buried oxide (BOX) is typically a few hundred nanometers thick, and the SOI is about 200-nm thick and is typically thinned down further (if desired) by oxidation.

BESOI is achieved based on the fact that very smooth and clean silicon and oxide surfaces will bond together when forced into intimate contact. When subsequently heated, the bond becomes permanent. One BESOI process involves bonding (to a handle wafer) a wafer with an epitaxial silicon layer, including a buried heavily boron doped etch stop. After bonding to the handle wafer, the back of this wafer is ground and then etched off down to the etch stop. Another BESOI process, the Smart-cut process, is much easier to scale up to large production volumes. For this process hydrogen is implanted under the surface of one wafer. After bonding, the wafer is heated which reacts the implanted hydrogen forming buried gas-filled cavities which join together, splitting off the back of the wafer and leaving the thin SOI layer still bonded. For some of the more advanced structures we will discuss below, BESOI is essential because it allows one to integrate a complex layered structure, patterned if necessary, onto the SOI wafer.

An alternate way of making SOI is to selectively grow silicon epitaxially from a local area (the seed) and have it grown over an adjacent isolation oxide [72] or even through predefined regions or tunnels [73]–[80]. An essential adjunct to the bonding and epitaxial processes are chemical-mechanical polishing (CMP) processes to resmooth the surface.

2) Partially Depleted (PD) SOI: PD SOI [81] is very similar to bulk CMOS with the SOI island taking the place of the n- or p-wells. The SOI is thick enough that the channel counter doping forms a conducting "body" under

TSMC Exhibit 1029

Table 4
SOI Variants

Variant	Strengths	Weaknesses
Partially depleted SOI	1. Channel design bulk-like 2. VT insensitive to BOX interface	1. Very susceptible to floating body effects (but solutions are available) 2. Same scaling constraints as bulk
Fully depleted SOI	1. Elimination of floating body effects 2. Elimination of punch-through currents 3. Elimination of drain-body tunneling	1. VT sensitive to SOI thickness and back interface 2. Back-channel potential may be influenced by drain voltage 3. Difficulty of contacting thin SOI
Ground Plane (GP)	1. Same as FD SOI 2. GP shields channel from drain 3. GP permits electrical control of VT 4. GP may be used as second gate	1. VT sensitive to SOI thickness 2. Difficulty of contacting thin SOI 3. Degradation of subthreshold slope by close GP
Double Gate (DG)	1. Maximum electrostatic control of channel and best scaling potential 2. Best current drive and performance 3. OR logic function within single device	1. Difficult to fabricate 2. Mis-aligned top and bottom gates result in extra capacitance and loss of current drive 3. VT control difficult by conventional means
Stacked SOI (ST)	1. High functional density 2. Shorter wires therefore higher performance and lower power	1. Fabrication complexity 2. Difficult to cool

the FET channel, whereas the source and drain implants usually penetrate to the back interface. Unlike bulk CMOS, the conducting body regions are usually left floating (they can be tied together using a separate body contact [82] or by leaving the islands connected by a continuous SOI region [83]). The floating body can charge up, causing dc effects such as premature breakdown and enhanced subthreshold slope at high drain bias, as is shown in Fig. 10. It also causes transient effects on circuit performance [84], [85], some good and some bad. On balance, the floating-body effects (FBE) are not desirable because they lead to irreproducible circuit performance. Body charge may be generated either by capacitive coupling to gate and drain, or by impact ionization of the drain current, and is drained away by recombination at the body/source depletion region. Promising solutions to the FBE involve increasing the recombination in the source by implanting to create recombination centers [86] or lower the bandgap [87].

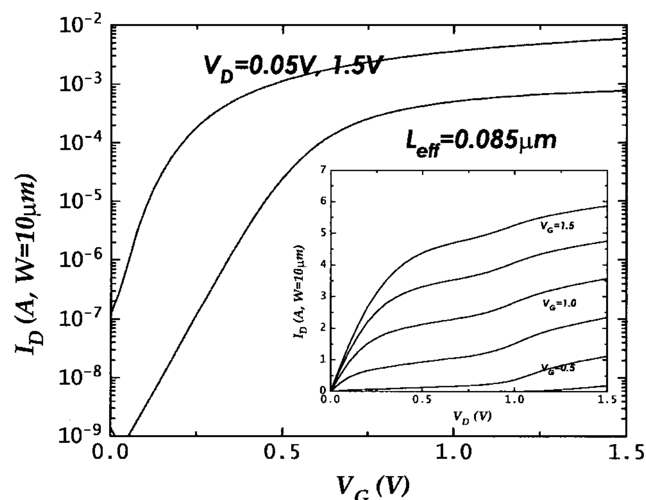
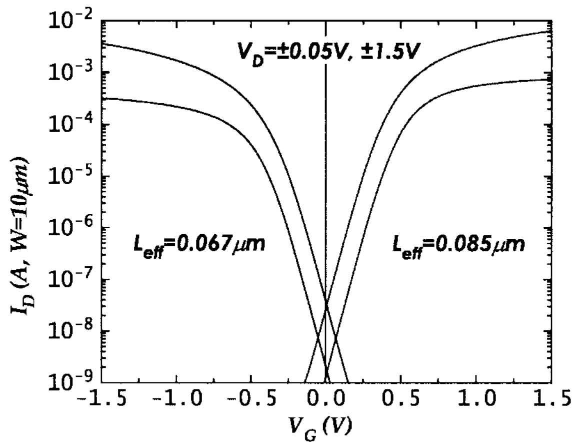


Fig. 10. I_D versus V_G and V_D characteristics of a 0.085-μm PD SOI NFET showing the floating-body effect, from Wann *et al.* [93].

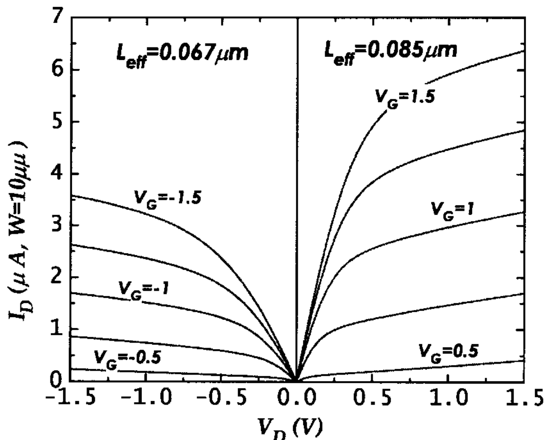
PD SOI, the forerunner perhaps of the more exploratory types, is on the threshold of commercial success with a range of impressive VLSI systems already having been demonstrated, including 4-Mbit SRAM [88], 1-Gbit DRAM [89], and high-end microprocessors [90]. In all cases except one [91], SOI demonstrated a significant advantage (from 15 to 50%) in speed and power over bulk CMOS. In Chau's case [91] the zero V_D threshold voltages had to be increased by about 0.2 V to overcome the decrease at operating V_D caused by the FBE, and this greatly impacted performance. The others [88], [90] have implemented solutions for the FBE and are not penalized as heavily. PD SOI has been scaled successfully to 70-nm gate length and record performance has been obtained [92], [93]. Subthreshold characteristics of these FET's are shown in Fig. 11 which closely resemble bulk characteristics without any signs of FBE up to the working drain voltage.

3) Fully Depleted (FD) SOI: FD SOI has a thinner and/or more lightly doped SOI layer than PD SOI so that there is normally a negligible concentration of holes/electrons in an N/P channel FET. This all but eliminates the FBE except at large enough drain voltages where hole/electron generation by impact ionization is important. A second supposed benefit of FD SOI is that the thin SOI confines the carriers better, resulting in higher

TSMC Exhibit 1029



(a)



(b)

Fig. 11. (a) I_D versus V_G and (b) I_D versus V_D characteristics of a 0.085-μm NFET and a 0.0-μm PFET built on PD SOI, from Wann *et al.* [93].

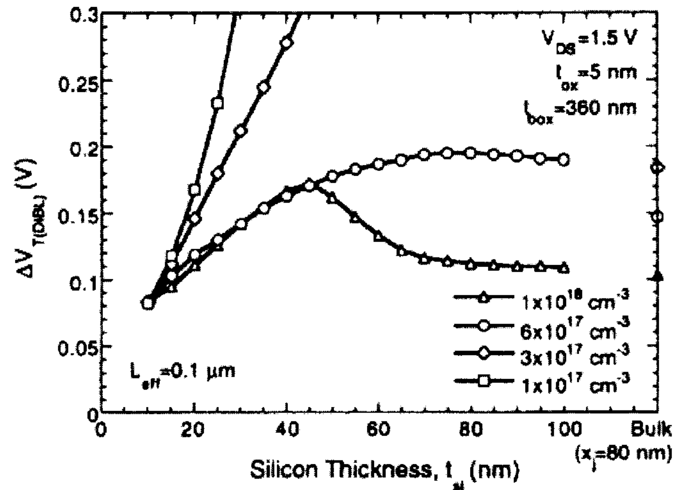


Fig. 12. Threshold voltage shift due to drain-induced barrier lowering as a function of silicon thickness for various channel dopings from Su *et al.* [94].

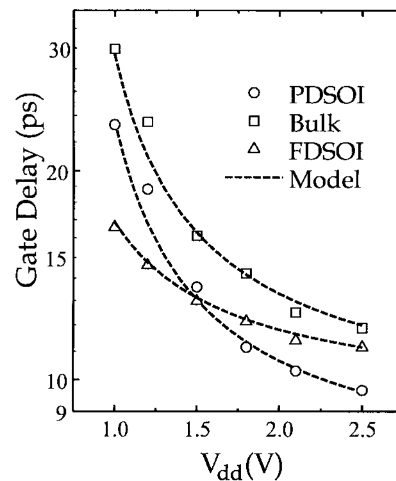


Fig. 13. Inverter ring oscillator delays comparing bulk, PDSOI, and FDSOI from Wann *et al.* [93].

mobility, better short channel effects, and better scaling potential.

This latter point has been investigated by several investigators [94]–[96], and using different modeling approaches, and experimentally by Wann *et al.* [93], [97] who show that, in general, FD SOI has a poorer scaling potential than bulk because of the lack of screening from the back of the channel. As illustrated in Fig. 12, FD SOI is predicted to have a much stronger drain induced barrier lowering than PD SOI for all but the thinnest (<40 nm) SOI layers. An inherent problem of FD SOI is the dependence of threshold voltages on the SOI thickness and the boundary conditions at the back SOI interface. Thus, while experimental VLSI circuits have been made in FD SOI [98], a large parameter spread has been noted.

Performance of FD versus PD SOI ring oscillator circuits has been compared by Wann *et al.* [93] in Fig. 13. These FD SOI transistors suffered from high series resistance (due to the thin SOI film, consumption of silicon by the self-aligned silicide process [99], and the lack of a thick source/drain fan out) which impacted their performance at high drain bias. In general, the contact technology is more demanding for the thin FD SOI transistors, but technologies such as the raised source/drain approach are being investigated [80], [100] which should alleviate this problem.

Mobility in thin SOI layers has been investigated both theoretically and experimentally. Theoretically [101], [102], mobility is expected to increase due to confinement and strain effects, but it should decrease due to additional interface roughness scattering from the back interface. Experimentally [103], [104], the SOI mobility (in the absence of strain) follows the bulk universal mobility model [105], [106] as is shown in Fig. 14, down to thicknesses of about 15 nm. The study in [103] showed mobility plunging steeply for $t_{Si} < 10$ nm, but this has yet to be confirmed by others.

The strain effects generally increase mobilities [107] especially for holes. In SOI, compressive strain is generated when LOCOS is used as isolation because of the volume difference between SiO_2 and silicon, and this has been shown experimentally [108] to cause up to 50% increase in hole mobility for SOI island widths of 1 μm.

TSMC Exhibit 1029

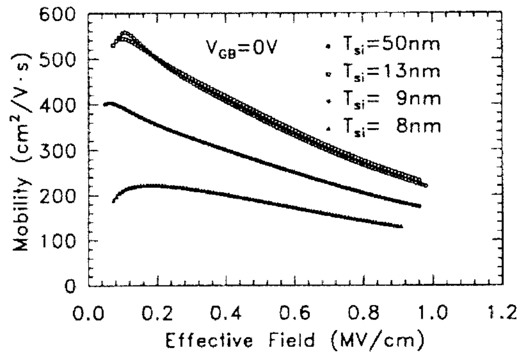
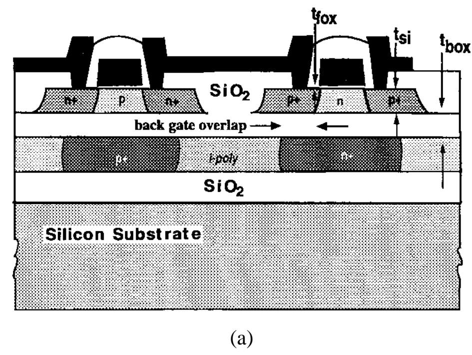


Fig. 14. Electron mobility versus effective field for SOI of various thicknesses, from Choi *et al.* [103].

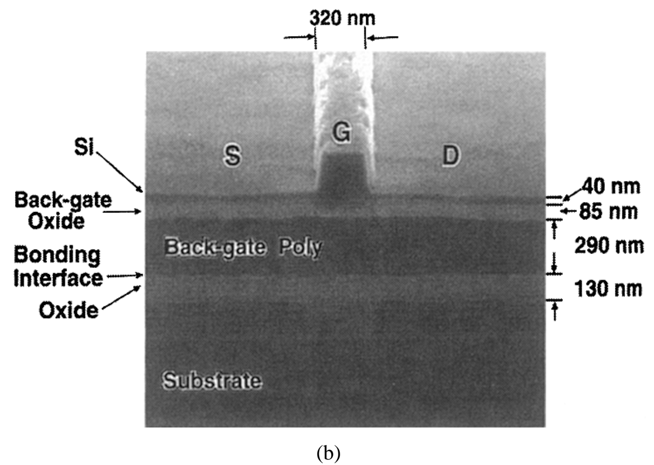
4) Ground Plane FET: The SOI structure can easily be extended to include a conducting layer underneath the silicon layer [104], [109]. This is called a "ground plane." Such an arrangement, as practiced by Yang *et al.*, is shown in Fig. 15. The ground plane may also be a doped well in the substrate itself [110]. The ground plane may serve two purposes. First, if the ground plane is close to the silicon channel, it screens the channel from the bottom against penetration of the drain field into the source and thus facilitates scaling to shorter channel lengths. In a way, it serves the role of the retrograde doping in a bulk MOSFET (by terminating drain electric fields) except that an insulating layer eliminates the source/drain-to-body p/n junction. This removes one of the limitations to scaling of the bulk FET, which is band-to-band tunneling current at that junction. To provide effective screening, the back gate insulator should be very thin, not more than about twice the thickness of the front gate oxide. The backgate insulator increases the "body effect," i.e., the shift in threshold voltage resulting from the channel to ground-plane voltage, and it also increases the subthreshold slope factor due to the capacitive division effect, $C_g/(C_g + C_{sub})$, between the gate and the substrate [111]. Unlike the bulk substrate, however, the ground plane can be partitioned and locally connected to the source to avoid the body effect.

Another use for the ground plane is as a means of shifting the threshold voltage of the top gate. The top gate threshold voltage may be controlled over the range between strong accumulation and strong inversion of the back interface, as is shown in Fig. 16, for the SOIAS structure [104]. For this purpose, the bottom insulator may be much thicker with a much smaller capacitive division ratio. As discussed by Yang *et al.* [104], power savings result from this approach since the threshold voltage of the transistor may be increased, reducing standby power, when the circuit is idle, and it may be reduced when the circuit is active to increase performance.

For both applications it is necessary to reduce the parasitic area of the ground plane, i.e., the area under the source/drain regions, to reduce parasitic capacitances which otherwise would reduce performance and increase power dissipation. Yang *et al.* [104] achieved this by selectively implanting the ground plane prior to defining the gate (nonself-aligned to the gate), while Horiuchi *et al.* [109] self-aligned the ground-plane to the gate by using a compensating implant under the source/drain (S/D) regions using the top gate as the self-aligned implant mask.



(a)



(b)

Fig. 15. (a) SOIAS back-gated CMOS device schematic and (b) SEM photograph of SOIAS cross section from Yang *et al.* [104].

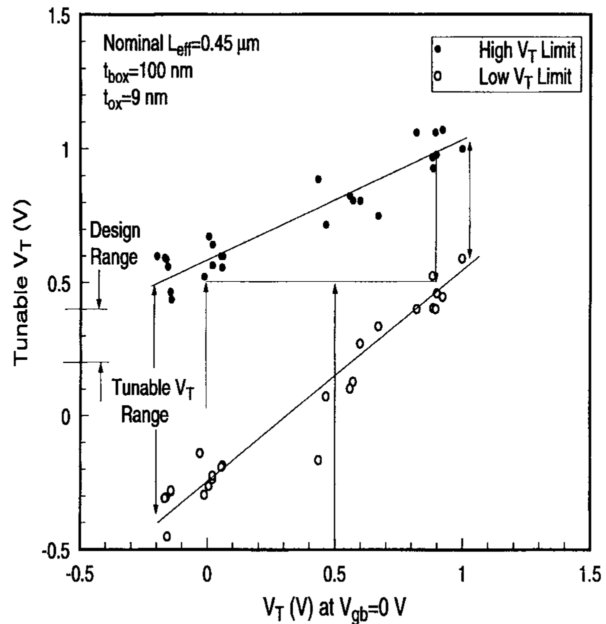
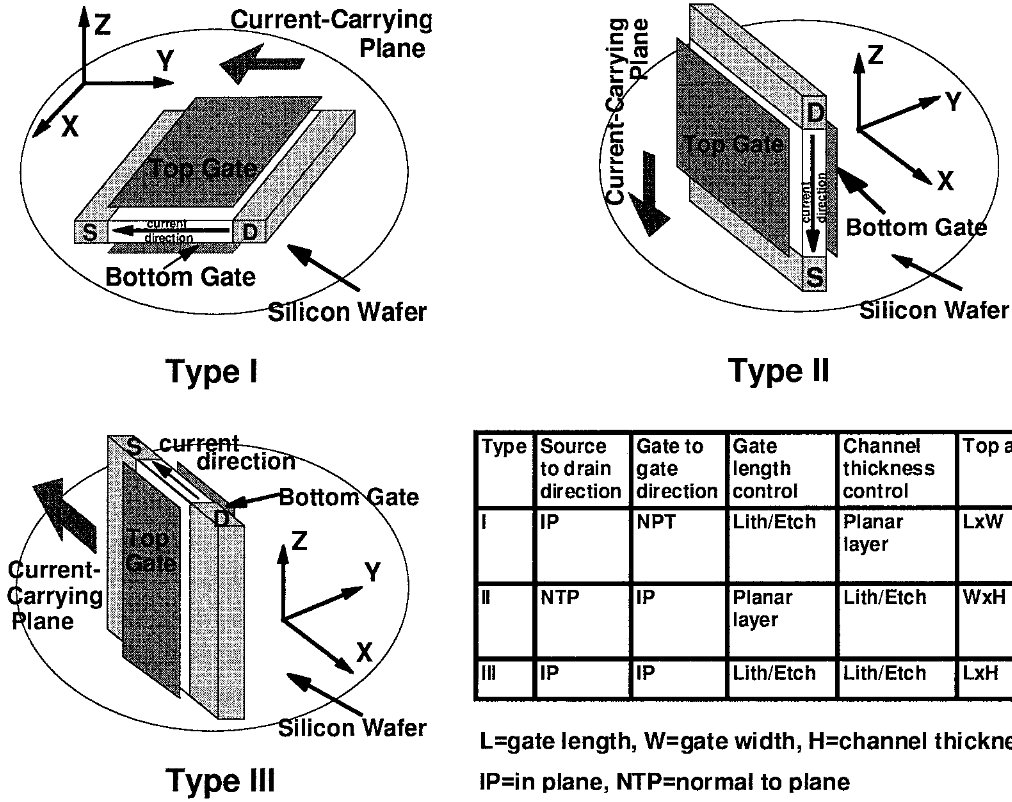


Fig. 16. Tunable threshold voltage V_T range via back gate biasing, as limited by the back interface becoming accumulated (low V_T) or inverted (high V_T limit). After Yang *et al.* [104].

TSMC Exhibit 1029

Table 5
DG FET—Topological Considerations (Adapted from Wong *et al.* [113])



Type	Source to drain direction	Gate to gate direction	Gate length control	Channel thickness control	Top area
I	IP	NPT	Lith/Etch	Planar layer	LxW
II	NTP	IP	Planar layer	Lith/Etch	WxH
III	IP	IP	Lith/Etch	Lith/Etch	LxH

L=gate length, W=gate width, H=channel thickness, IP=in plane, NTP=normal to plane

With the ground-plane confined under the (top) gate, one may question how this structure is different from that of a DG FET [39] (described in Section III-A5). The answer is that there is little formal difference except that in the ground-plane concept the bottom gate is inactive during ordinary circuit operation so that it may have a larger resistance without unduly impacting circuit performance.

5) Double-Gate (DG) FET's: The DG FET is electrostatically much more robust than the standard single-gated MOSFET since the gate shields the channel from both sides, suppressing penetration of the field from the gate, reducing short-channel effects [39]. For conventional, single-gated FET's the substrate plays the part of the bottom shield; yet this results in a tradeoff, as pointed out by Nowak *et al.* [112], between the degree of shielding and the reduction of the subthreshold slope, as discussed above. In the DG FET this tradeoff does not exist, and both gates are strongly coupled to the channel to increase transconductance. The relative scaling advantage of the DG FET is about two times [39]. The performance of the symmetrical version of the DG FET is further increased by higher channel mobility compared to a bulk FET since the average electric field in the channel is lower, which reduces interface roughness scattering according to the universal mobility model [105], [106].

The DG FET can be (and has been) made in three basic configurations labeled types I, II, and III in Table 5 [113]. Type I has the advantage that the channel layer is in the plane of the silicon wafer surface so that the channel thickness, the most critical dimension in the FET, is controlled by the thickness of uniform planar layers rather than by lithography. The type II DG FET, which has the channel in the vertical direction, is most compact for DRAM application [114] where low leakage current (hence a long channel) is important and performance is secondary, but it has obvious topological difficulties for a CMOS logic application. The type III DG FET, while facing formidable technological difficulties for implementation at minimum dimension, has the highest packing density for high-speed logic applications since the channel width, the longest dimension of a logic FET, is perpendicular to the plane of the wafer; furthermore, all four terminals of the FET are accessible from the top. A version of this has been implemented in [115]. For the type II DG FET's, and especially for the type III DG FET's, a technological solution would have to be found for the otherwise crippling dependence of the critical channel thickness parameter on lithography and etching acuity. Possible material solutions exist, such as using sidewalls for controlled thickness and using crystallographic selective etches to form a channel on the perpendicular (1, 0, 0) etched plane, but work on these approaches [116] is still at a primitive stage.

Although the various configurations of Table 5 emphasize the importance of the channel, it should be noted that apart from a thin (<10 nm) silicon channel, a high-performance DG MOSFET must also have a thick source/drain fan out to reduce the series resistance, as well as a set of gates that are self aligned to each other and to the source/drain doping regions [113], [117].

TSMC Exhibit 1029

Table 6
Double-Gate (DG) FET Electrical Family

Style	Description	Advantages	Disadvantages
SYMM	Intra-gap gates equally spaced from thin undoped channel	1. "Ideal" double-gate structure 2. Separate control of both gates 3. High channel mobility 4. No dopant fluctuations	1. Inflexible VT control 2. Channel doping leads to severe VT fluctuations 3. Quantum effects limit minimum channel thickness 4. Space needed for extra gate contact
N+/P+	High-low gate work-function combination to simulate mid-gap gate work-function	1. Can use CMOS standard n+/p+ polysilicon gates 2. Better electron confinement relaxes silicon thickness requirements	1. Lower mobility than symmetrical structure 2. Reduced control of gate furthest from channel
WRAP	Gate wraps around silicon beam • Narrow beam plus undoped channel • Wide beam plus doped channel	1. Very compact structure (wide beam) • Even better gate control than "ideal" structure (narrow beam) • Variable device width (wide beam)	1. Crystal orientation varies around channel (narrow beam) 2. Interface states and VT smearing 3. Source/drain fan-out difficult to fabricate 4. Small current carrying capacity per channel; inflexible VT control; enhanced quantum effects (narrow beam) 5. VT very sensitive to beam width; difficult doping profile control; wider structures have poorer scaling properties (wide beam)
DTMOS	Connected body acts as a back junction FET	1. Closest to standard SOI CMOS implementation 2. Flexible VT control	1. P/N junction turn-on limits gate voltage 2. High resistance of body gate

Different types of DG FET may also be classified according to the electrical function of the different layers present and how they control the dimensionality and electric field configuration in the channel. Such a representation is given in Table 6 and may be applied, in general, to any of the configurations in Table 5.

6) Symmetrical DG FET: The symmetrical DG FET is the purest implementation of the DG FET. To obtain suitable threshold voltages for logic, the gate material should have approximately a mid-gap work function. As such, the gate material would be suitable for both p- and n-channel FET's. Such an FET, with an undoped channel, does not suffer from V_T variations caused by random dopant fluctuations in the channel. However, a single gate work function results in a fixed threshold voltage ($\simeq \pm 0.4$ V) which may be too rigid for most applications.

In principle, one could fine tune the threshold voltage by ion implantation as in bulk FET's (the dopant may be of either polarity depending on the sign of the desired shift). This will reintroduce doping fluctuations which may be larger than in bulk FET's since the dopant has to be placed in the channel itself where the effect of fluctuations is maximized, whereas in the bulk a retrograde doping profile can be used (see Section II-D3). On the other hand, fluctuations will be reduced by the fact that the gate capacitance is effectively doubled, and for mid-gap gates the V_T only needs to be shifted a small amount (0.2 V versus 0.7 V for bulk). An intriguing possibility for V_T adjustment is to change the work function of the gate. This is feasible if one gate is an alloyed semiconductor such as $\mathrm{Si}_{1-x}\mathrm{Ge}_x$ or $\mathrm{Si}_{1-x}\mathrm{C}_x$. While preliminary studies have been made [118], [119], it is still not at all clear how practical this approach will be.

When the channel becomes very thin, V_T will vary because quantum shifts of the ground state energy is inversely proportional to the square of the silicon thickness. In silicon, this shift in smallest for a (1, 0, 0) channel. Assuming a simple particle-in-a-box model [39], [120], the uncertainty of the threshold voltage (σ_{V_T}) is $\sigma_{V_T} = -(\hbar^2\pi^2/qm^* t_{\mathrm{Si}}^2)(\sigma_{t_{\mathrm{Si}}}/t_{\mathrm{Si}})$. For a 4-nm thick silicon channel with a 20% channel thickness control $(\sigma_{t_{\mathrm{Si}}}/t_{\mathrm{Si}} = 0.2)$, the σ_{V_T} is 50 mV, which is too high (see Table 2 and Fig. 6). This is illustrated in Fig. 17, where the V_T varies rapidly as the channel thickness goes below 5 nm [121]–[123].

Examples of symmetrical FET's are shown by Tanaka *et al.* [124], Denton *et al.* [125], and by Wong *et al.* [113], using either P^+ polysilicon (Tanaka, Denton) or N^+ (Wong) polysilicon gates. The techniques used were BESOI (Tanaka) similar to the SOIAS structure discussed in Section III-A4, and with a much thinner, seeded, unconstrained epitaxial growth (Denton), and constrained epitaxial growth through a thin tunnel (Wong). The basic structure for Wong's transistor is shown in Fig. 18. Also, shown in Fig. 19 is the amazing ability of the epitaxial growth to thread through thin tunnels. CMP was used in all cases for planarization. In the first two cases (see Fig. 20) the top and bottom gate were not self aligned. Only in Wong's case was the top and bottom gate self-aligned and device symmetrical with respect to oxide thickness. Wong *et al.*

TSMC Exhibit 1029

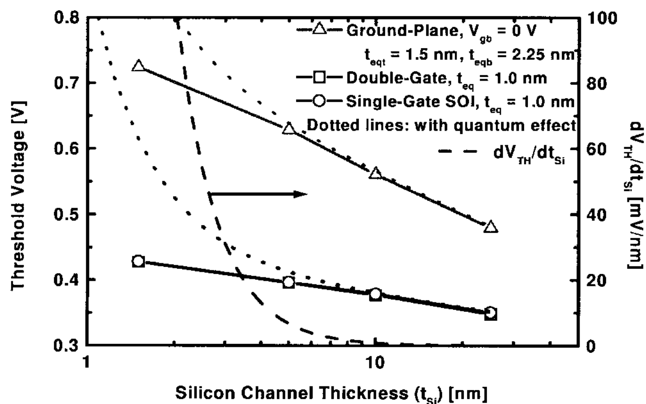


Fig. 17. Left axis: Dependence of the threshold voltage (long channel, without quantum mechanical effects) on the silicon channel thickness for DG, ground-plane, and ultrathin fully depleted SOI MOSFET's. The estimated additional shift of the threshold voltage due to quantum confinement of the thin silicon channel is shown as dotted lines. Right axis: the estimated sensitivity of the threshold voltage to the silicon channel thickness (dV_{TH}/dt_{Si}) due to quantum effect. DG and single-gated SOI: gate oxide = 1 nm. Ground-plane: top gate oxide = 1.5 nm, bottom gate oxide = 2.25 nm. After Wong *et al.* [123].

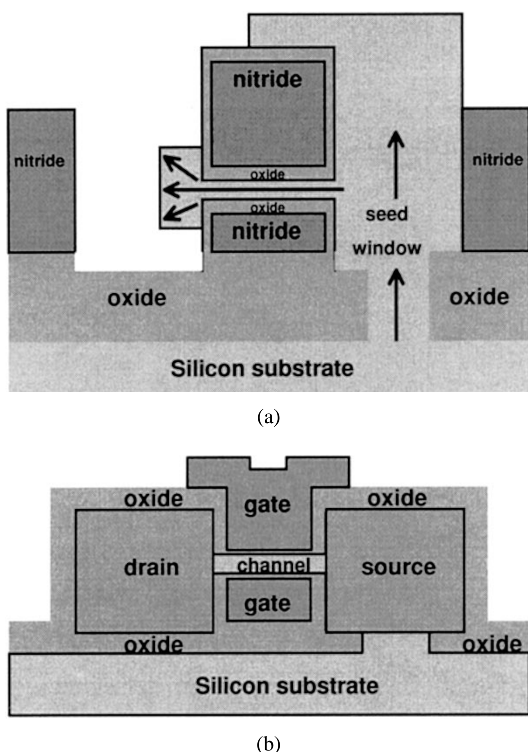


Fig. 18. Schematic illustration of (a) selective epitaxial silicon growth through a thin tunnel and (b) cross section of completed DG MOSFET. After Wong *et al.* [113].

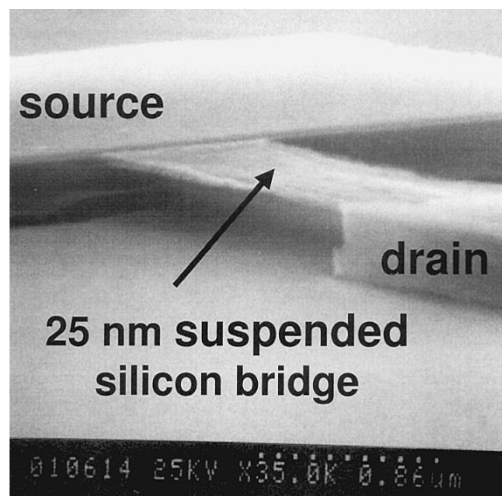


Fig. 19. SEM image of a 25-nm suspended silicon bridge. Growth of gate oxide and conformal deposition of the gate material completes the process. After Wong *et al.* [113].

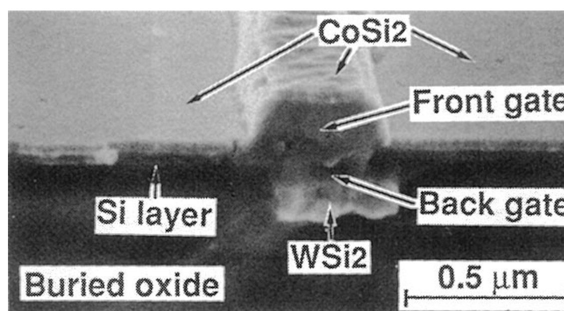


Fig. 20. SEM image of a misaligned P^{+}-N^{+} DG MOSFET, from Tanaka *et al.* [126].

analyzed the impact of misalignment on performance [117], where a 25% misalignment caused a 33% increase in the delay, thus highlighting the importance of alignment of the gates.

7) P^{+}-N^{+} Gates: To effectively get a mid-gap work function using conventional polysilicon gates, Tanaka *et al.* [126] used top and bottom gates of P^{+} and N^{+} polysilicon. This results in an asymmetrical FET with only a surface channel closest to the gate of like polarity being turned on for low gate bias, the threshold for the other channel being approximately 1 V higher. The field distribution across the channel is asymmetrical, much like in a bulk FET, and the channel is better confined than for an equivalent symmetrical FET. Tanaka's FET's were made using the same process as his P^{+} gate FET's mentioned above. They are among the few high-speed results published for DG FET's, and their performance (for 0.19-μm gate length FET's) is a very respectable 43 ps at $V_{DD} = 1$ V at room temperature, going down to 15 ps at 77K [127].

8) Wrap-Around Gate: In the wrap-around gate FET [also called surround gate [128] or gate all around (GAA) [129]–[131] in some implementations] the gate wraps around a beam of silicon, which may be wide or narrow, rectangular, square, or circular, and of any of the orientations of Table 5. Unlike conventional MOSFET's, the channel is not a (100) silicon interface (which has the lowest interface state density) but spans a range of orientations. This has led to kinks and other structure in the subthreshold characteristics in some structures [132], but others [114] seem well behaved. The implications of this for performance of logic devices has yet to be assessed.

Structures with large cross sections and doped beams behave like bulk FET's [114] and will not be further discussed. The most interesting structures have very narrow

TSMC Exhibit 1029

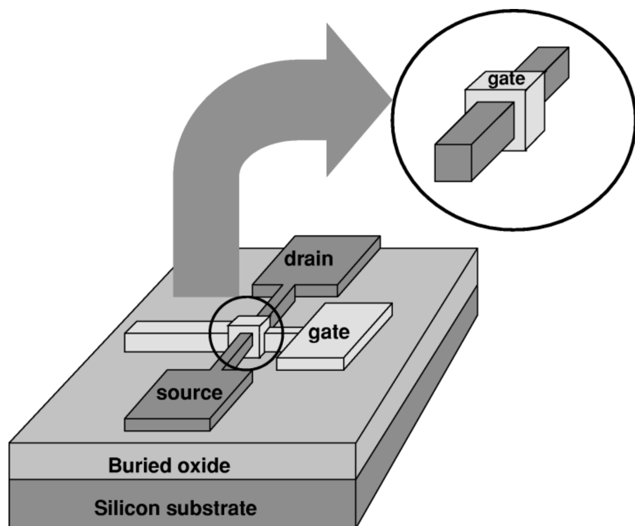


Fig. 21. GAA FET, schematic structure showing current flow along all four surfaces, adapted from Leobandung *et al.* [131].

beams, where the electrostatic influence of the gate on the channel is felt from all sides [131]. The structure of Colinge [129] used hydrofluoric acid to undercut an oxide supported beam. A subsequent conformal polysilicon deposition forms the surrounding gate. This is shown schematically in Fig. 21.

Studies [128] show that this extra degree of electrostatic confinement permits the GAA FET to be scaled about 50% further than the DG FET, and this trend is confirmed experimentally by Leobandung *et al.* [131]. This property may also be used to relax the thickness requirements on the silicon and oxide layers for a given gate length. This scaling advantage is offset by the fact that, for an isotropic effective mass, the V_T shift due to quantum confinement is larger for a cylindrical enclosed structure than for a the planar structure of thickness equal to the cylinder diameter by 2.3 times, and even more compared to a (100) surface, due to the anisotropic effective mass. On balance, therefore, it appears that the argument for better scalability is questionable.

The main advantage of the GAA approach is extremely high packing density for the case of vertical structures (Type II) which is good for memory, whereas the main disadvantage is the very limited current-carrying capability per device which is a severe handicap for high-speed logic applications where large wiring capacitances have to be switched. GAA in the Type III configuration does not offer higher packing density or higher current per unit planar area because each "wire" or "beam" requires one lithography line and space, and there is little area advantage unless an unusually large vertical height is used. Silicon channels with large vertical height and uniform thickness are difficult to achieve by pattern and etch processes.

9) Dynamic Threshold MOS Transistor (DTMOS): DTMOS [133] is the poor man's DG FET, where the doped body, in the context of PD SOI, acts as a back, p/n junction gate. Its range of applicability is limited by the turn-on voltage of the p/n junction. The back gate may also be driven by an additional transistor [134] to maintain a high input impedance. The individual FET's have to be very narrow in order to propagate high-speed signals down the highly resistive back gate [97]. The failure to do so leads to deleterious effects of increased delay and, more seriously, increased power dissipation due to the delayed turn off of the back gate. For instance, for a body resistance of 20 KΩ/sq (typical of PD SOI), a total capacitance per unit width of 2 fF/um, and a delay of 1 ns, the maximum width per FET finger is 2.8 μm. Overall, this approach may be useful for low power, low voltage circuits where the above restrictions are not too severe.

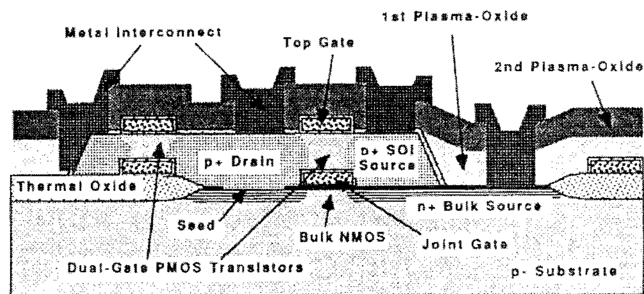


Fig. 22. Schematic cross section of a stacked CMOS inverter made with selective epitaxy. Note that the gate of the bulk NMOS device also acts as a bottom gate for the DG PMOS device. After Zingg *et al.* [135].

10) Beyond DG FET—Stacked FET's: With the ability to do repeated bonding and polishing as in Tanaka [124], [126] or repeated epitaxial growth and polish back as in [125], it becomes possible to stack multiple layers of devices on top of each other, as demonstrated by Zingg [135] (see Fig. 22) where a CMOS inverter was built with DG p-channel devices stacked on top of n-channel devices with the (top) gate of the n-FET serving as the bottom gate of the p-FET. Techniques such as these may be useful in the future for instance to make dense memory arrays on a logic chip, although it is questionable whether the logic chip will benefit as much because of the issue of via blockage in connecting the multilayer structure [136].

The main problem with the above approaches is that the layers are processed serially, which means that the increased density does not result in cost savings (indeed, cost may increase superlinearly due to the difficulty of yielding the multiple layers).

Today's IC's involve many wiring levels (e.g., IBM's six-level copper-damascene process [28]), and in the future this will increase. A revolutionary strategy would be to integrate the devices into the interconnects process, so that device and wiring is built up together. We see in the techniques used above for the various types of devices perhaps the first tentative steps in this direction.

11) The Problem of Heat Dissipation: In SOI and especially for stacked devices, heat dissipation is potentially a problem because the thermal conductivity of silicon dioxide is much poorer (100 times) than silicon. This has been a problem, especially in the measurement of I–V characteristics of SOI devices where large gate and drain voltages

TSMC Exhibit 1029

Table 7
Analysis of Self Heating in SOI (Assumptions Are $t_{si} = 100$ nm, $V_G = V_{DD}$, $V_T = 0.25$ V, Interdigitated FET's with Period of 6 × the Effective Channel Length, AC = 5% Duty Factor)

	$V_{DD}(V)$	Current (A/cm)	Power Density (kW/cm^2)	DC Temperature rise (K)	AC Temperature rise (K)
0.1 μm single gate FET	1.8	7	210	150	7.5
0.05 μm double gate FET	1.2	10	400	286	14
0.03 μm double gate FET [39]	0.9	15	750	536	27

are maintained under essentially dc conditions. For a single device, the temperature rise may easily exceed 100°C [138], and DG devices may reach even higher temperatures due to their larger current-carrying capability. Jenkins and Sun [137] used high-speed measurements with 7-ns pulses to circumvent this problem, utilizing the property that the thermal time constants (tens to hundreds of nanoseconds) are longer than the measurement pulses.

Local temperature rise due to thermal resistance of the buried oxide is estimated in Table 7 for several SOI transistors. A multifingered FET (typical of a driver FET) is assumed. This is in a sense the worst case because a large device width is crammed into a small area. The temperature is assumed to be uniform over the area of the SOI island, which is a reasonable assumption, given the large thermal conductivity of silicon, even for the thinnest SOI layers (10 nm) likely to be encountered in practice. Heat first flows laterally, heating up the entire silicon island, then it flows vertically, resulting in a temperature drop across the BOX. Table 7 shows the temperature rise for devices chosen to represent: 1) today's state of art [92]; 2) a future generation DG FET; and 3) the most severe (highest current) predictions for a future "ultimate" DG FET [39]. The BOX thickness of 100 nm was chosen to minimize temperature rise while still not degrading performance [an even thinner BOX layer could be used to advantage for cases 2) and 3)]. While it would not be advisable to measure the DG FET's under dc conditions (unless extra heat-sinking capability were provided via the wires [138]), the ac temperature rise is quite comfortable. For ac, a fairly high duty factor of 5% was chosen (see Section II-D4) but even this might be exceeded on a heavily loaded (i.e., large risetime) driver showing that thermal considerations will pose additional constraints on circuit design.

For the chip as a whole, the situation is more complicated. On the one hand, the average power densities, even for a high-power processor chip (e.g., 100 W for a 1 cm^2 chip), are very small compared to local power densities discussed above. This leads to a very small penalty imposed by SOI since the average thermal resistance of the BOX (using the total SOI area) is substantially less than that of the silicon substrate (the substrate being much thicker than the BOX) and these thermal resistances are much less than that of a typical silicon package, by more than an order of magnitude. On the other hand, the BOX does exacerbate the temperature rise of hot spots, due to locally high activity on the chip. Mainly small hot spots are affected, of the order of 10 μm or less (about 100 times the thickness of the BOX) because the thermal resistance of smaller spots is dominated by the BOX, whereas that of larger spots is dominated by the bulk silicon.

For the more aggressive device packing schemes shown in Table 5 (Types II and III), and for the stacked structures, the thermal constraints are more severe but do not appear to be insurmountable given the overall trends toward lower power and the possibility of distributing this heat along the wires.

B. Memory

For use as a memory element, the logical compliment to nanometer scale MOS logic devices would be an floating-gate (FG) transistor, or FLASH memory (Fig. 23). Being single-element devices, they can be densely packed and integrated relatively easily with logic transistors. Since they consist of a standard transistor with an additional FG layer in the gate stack, some of the same scaling considerations apply to these devices as to logic devices with the exception of retention time. There are, however, a few considerations unique to these structures, as discussed below.

1) Structure: Fig. 23(a) shows a cross section of a generic FLASH memory device, showing the floating island of silicon embedded in the gate oxide. Memory is achieved by tunneling carriers on and off this island, shifting the threshold of the underlying channel region. If the size of the island is scaled to small dimensions (on the order of 5 nm), the increased electron energy level spacing due to quantum confinement, combined with the Coulomb blockade effect, can be used to precisely control the charge on the island to one (or a few) electrons at room temperature. In this size range, the floating island effectively becomes a quantum dot electrostatically coupled to the transistor channel. This allows one to take advantage of confinement for controlled writing and erasing of the memory without having to accept the impedance mismatches and low currents inherent to single-electron tunneling in the channel (i.e., drive current) of the transistor.

The scaling of the island can be achieved in two ways. For small devices, where the width of the device is less than 10 nm, a single floating island self aligned to the channel can be used, as shown schematically in Fig. 23(b). This

TSMC Exhibit 1029

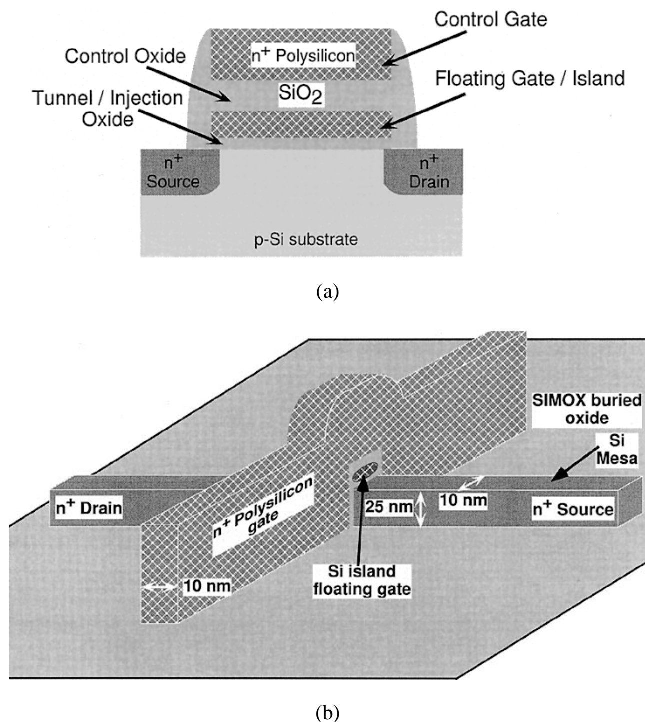


Fig. 23. Schematic of (a) a generic FLASH memory structure cross section and (b) a quantum-dot FLASH memory.

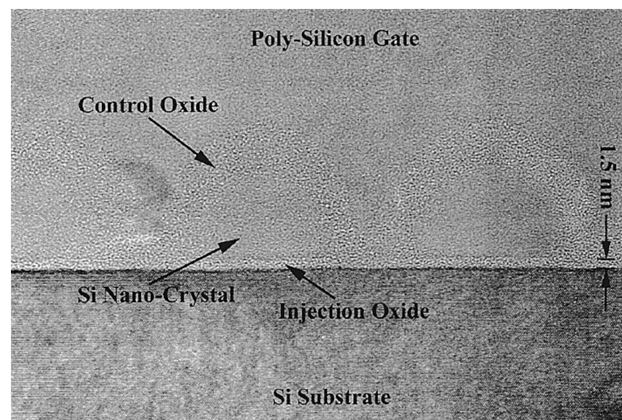


Fig. 24. TEM of a FLASH memory showing three silicon nanocrystals embedded in a control oxide on top of a thin tunnel oxide.

basic structure has been used by several groups [139]–[141] and both single-electron charging and memory effects have been realized at room temperature. However, this design limits the channel width to be within a Debye length of the dot size, limiting the current drive. For larger devices, an array of quantum dots, or nanocrystals, can be embedded in the gate oxide [142], as shown in the TEM in Fig. 24, to cover the whole channel width. This can be achieved by either direct deposition or implantation into the oxide followed by annealing, although the former is preferable to allow better control of the nanocrystal size, position, and characteristics.

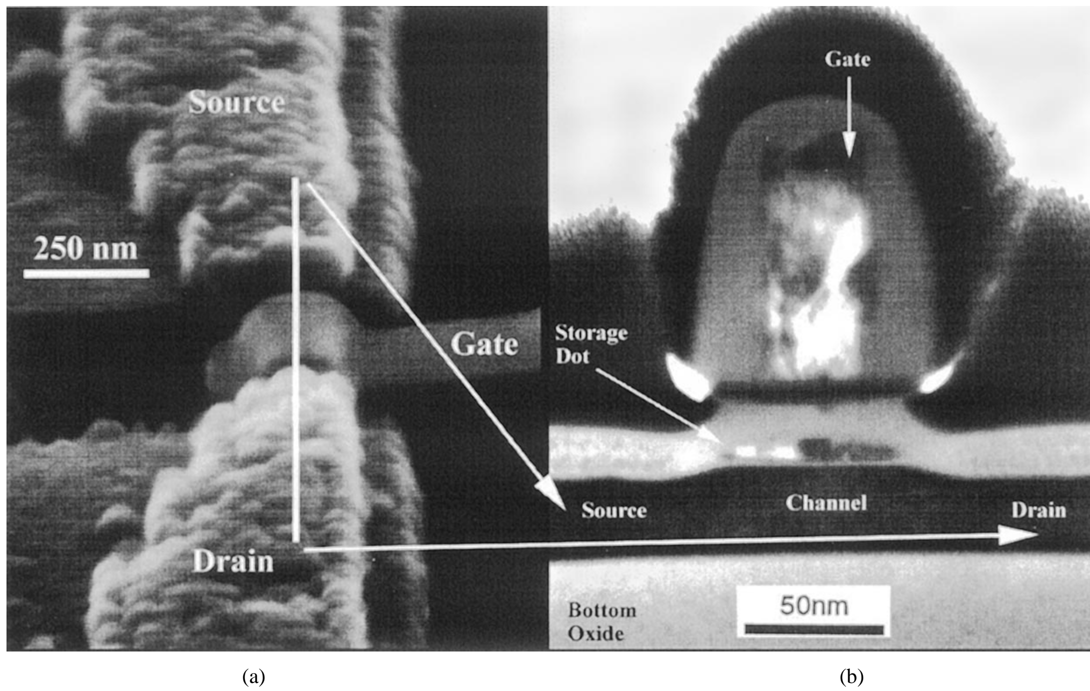
An alternative structure, in which the channel itself is fabricated in ultrathin polysilicon, has been demonstrated by Yano *et al.* [143]. In this case, the charge is thought to be stored in the polysilicon grains or grain boundaries, effectively shifting the threshold of the electron channel which winds its way through the adjoining grains. This structure has yielded some of the most interesting room temperature demonstrations to date, including a small (8 $\times$ 8-b) memory array [144], as well as a new prototype architecture aiming for densities of 100 Mb and beyond [145]. Current drawbacks of this structure include the high resistance of the channel, relatively high program voltages required, and the cell-to-cell variation. The last of these has recently been addressed using a new writing/sensing technique [146], which may also be applicable to quantum-dot memories in general. If the other drawbacks can be similarly mitigated, the structure may turn out to be a useful alternative to the more common "scaled-FLASH," single-dot memory transistor.

2) Scaling: Both the nanocrystal and single dot memory transistors can be fabricated almost identically to a logic transistor. The only changes are in the gate stack deposition, where the floating-gate layer must be included, and the subsequent gate stack etching. An example of a fabricated quantum dot memory is shown in Fig. 25 [141]. In this case, the device is fabricated on SIMOX, and the floating-gate island is unintentionally larger than the gate length. Ideally, the gate length would most likely be larger than the island. This length, along with the overall transistor channel design, would be dictated by the need to control short channel effects. Since any source-to-drain leakage currents would not affect the memory storage capability of the device (unlike a DRAM), the criteria for scaling the transistor will be the same as for the logic transistors, where a balance between drive current, off-state leakage, and overall power must be maintained.

The scaling of the gate stack, however, is unique to the memory devices. It is similar for both the multiple and single dot case, which can be seen as a progression of the same device. They will be discussed together below [147] under the generic label of quantum-dot memory. Using current lithographic techniques, it is very difficult to fabricate a single dot of small enough size to take advantage of single electron effects at room temperature, but nonlithographically formed nanocrystals embedded in the gate oxide of a larger transistor can already achieve these sizes. While uniformity of size, orientation, and charge state may ultimately limit the viability of using true single-electron charging in either single or multiple dot structures, many advantages in terms of power consumption and noise margin can still be gained by scaling the floating gate to as small a size as possible.

a) Storage dot size: While the maximum dot size is largely controlled by the desire to scale and control the number of electrons stored, the minimum size is limited by power and speed concerns. Table 8 lists various parameters for an ideal quantum-dot memory as a function of dot size. As the dot size decreases, the energy spacing (as measured by the ground state eigenenergy in the table) becomes

TSMC Exhibit 1029



(a) (b)

Fig. 25. (a) SEM and (b) TEM of a quantum-dot FLASH memory.

Table 8
Approximate Self-Capacitance, Charging Energy, and Ground State Eigenenergy Due to Quantum Confinement for a Silicon Sphere of Given Diameter Embedded in an Oxide (Assuming a Control Oxide Thickness of t_{CG} = 7 nm, Conformal to the Sphere, the Gate Capacitance and Resulting Threshold Voltage Shift Due to Single-Electron Storage Are also Computed)

Quantum-Dot Diameter, $2r$ (nm)	Capacitance, $C \approx 4\pi\varepsilon_{ox} r$ (aF)	Charging Energy, $e^2/2C$ (eV)	Ground State Eigenenergy (eV)	Control Gate Capacitance*, C_{CG} (aF)	Single Electron ΔV_T, e/C_{CG} (V)
30	6.51	0.012	≈0.003	4.10	0.039
20	4.34	0.018	≈0.007	2.11	0.076
10	2.17	0.037	≈0.030	0.63	0.255
5	1.08	0.074	≈0.104	0.17	0.924
3	0.65	0.123	≈0.290	0.07	2.460
2	0.43	0.185	≈0.600	0.03	> 5
1	0.22	0.369	≈2.600	0.01	> 20

$* \, C_{CG} \approx (2\pi\varepsilon_{ox} r^2) / (t_{CG} + (\varepsilon_{ox}/\varepsilon_{Si})\, r)$

large. This increases the amount of gate voltage needed to tunnel into the state, increasing the power requirements. In addition, the number of states available for tunneling in any given energy range decreases, resulting in lower efficiency and speed. Therefore, for efficient operation in the 1–1.5-V range, a dot size of 3–6 nm seems optimum.

b) Control oxide thickness: The minimum size here is constrained by the need to avoid leakage to the control gate, as well as between dots in the multiple dot case. Note that isolating multiple dots from each other has the advantage of making the structure more robust to defects in the oxide, since a point of high leakage will only cause a few dots to lose their charge. Similarly, only nanocrystals directly over the drain overlap region will be affected by the higher leakage expected from the accumulation layer there [37], or by bit-line disturb when other cells in an array are being accessed. The maximum control oxide thickness must be limited to reduce the gate voltage needed, making the optimum thickness on the order of 5–10 nm.

c) Injection oxide thickness: This is the most crucial parameter affecting the memory performance, as it has the most control over the currents into and out of the quantum dot. Simulations and experimental results (Fig. 5) have shown the exponential dependence of tunneling (i.e.,

TSMC Exhibit 1029

Table 9
Extrapolated Refresh Times from Measurements on Devices with Comparable Nanocrystal Density

Oxide Thickness	Write Conditions	Threshold Shift, ΔV_T	Refresh Time
16 Å	200 ns, 3 V	≈ 0.65 V	20°C > 1 wk 85°C ≈ 1 hr
21 Å	400 ns, 3 V	≈ 0.48 V	20°C > 1 wk 85°C ≈ 5 hr
30 Å	1 μs, 3 V	≈ 0.55 V	20°C > 1 wk 85°C >>1 hr
36 Å	5 μs, 4 V	≈ 0.50 V	20°C Large 85°C Large

write/erase) current on oxide thickness [35], [36]. In particular, a large increase in the current at low voltages is observed for oxides in the direct tunneling regime (<3 nm). Since large currents persist in these thin oxides down to arbitrarily low voltages, the tradeoff is between speed and volatility, as seen in Table 9.

From a fabrication standpoint, the thinnest oxide that can be produced reliably with integrity is about 1.5 nm. From the table, this could allow write times in the 100-ns range but would only have a retention time on the order of days at best (these data do not include any potential read/write disturb effects which would be present in a real memory array). In this case, a refresh circuit, similar to what is used in DRAM, would be necessary. To obtain nonvolatility, the oxide would need to be on the order of at least 4 nm, increasing the write time to the 10-μs range.

Finally, making the injection oxide excessively thick must be avoided not only for the sake of maintaining high speed at reasonable power levels, but also to maintain high endurance. Thin oxides which allow direct tunneling have shown endurance up to 10^9 write/erase cycles with negligible narrowing of the threshold shift window [148], [149], while standard FLASH memories with thicker oxides which rely on Fowler–Nordheim tunneling or channel hot electrons for programming can only be cycled up to 10^5 times. Clearly, the final decision on oxide thickness will depend on the intended application.

d) Number of electrons stored: This last parameter is actually controlled by a combination of the thicknesses and sizes in the gate stack and the operating voltage. Obviously, for high speed and low power, the minimum number of electrons is desirable, but this must be balanced by the need for robustness and the constraints of stray offset charge (e.g., interface states) which can also cause threshold shifts. Fig. 26 shows the results of some simulations on the effect of offset charge due to interface states on the quantum dots, as well as dopant/charge fluctuations in the channel depletion region. For a dot size of about 5 nm, to insure the threshold voltage shift due to stored electrons is significantly higher than due to the other fluctuations [147], and to allow for statistical charge variations in the dot [150] as well as the loss of at least one or two electrons due to leakage between refresh cycles, it appears that storing on the order of five electrons is sufficient.

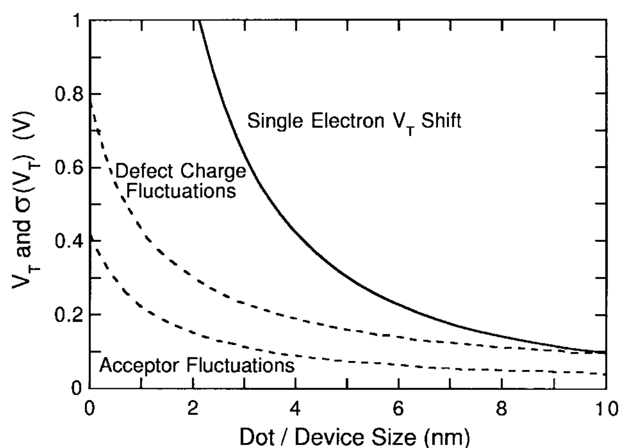


Fig. 26. Magnitude of the threshold voltage (V_T) shift due to a single electron in a 5-nm dot with a 5-nm control oxide, as compared to the standard deviation in V_T due to charged defects and acceptor fluctuations.

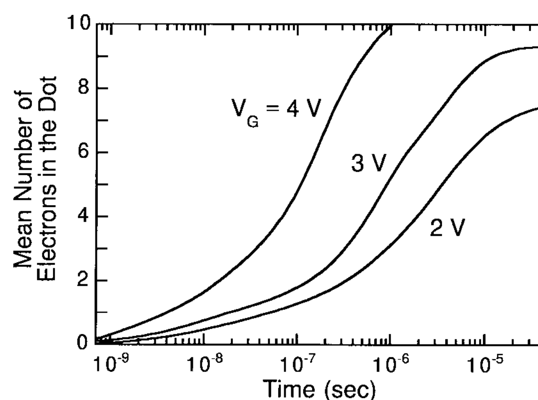


Fig. 27. Average number of electrons on a 6 × 10× 10-nm dot (t_{inj} = 1.5 nm) as a function of time during a write operation for varying gate voltage.

Note that for a device with multiple dots in the gate, the number of stored electrons per dot may be reduced, since the increased number of dots helps to mediate variations on any single dot.

3) Performance: Taking these scaling considerations into account, simulations were performed to evaluate the expected performance of these memory devices to help determine appropriate applications for their use [147]. Fig. 27 shows the mean number of electrons on the dot as a function of time for various gate voltages. (For this structure, each electron induces approximately 0.3 V threshold shift.) As expected, increasing the gate voltage increases the number of electrons the dot achieves in equilibrium and decreases the charging time. For all voltages, it requires on the order of 10 μs to achieve equilibrium, but if only a few electrons are needed, write times can be on the order of 100 ns.

Corresponding erase times are shown in Fig. 28. In this case, since the dot must always be restored to zero electrons, the erase time will be on the order of 10 μs for reasonable voltages, regardless of the initial number of electrons on the dot. It should be noted that most experimental results to date have actually shown significantly longer erase times than

TSMC Exhibit 1029

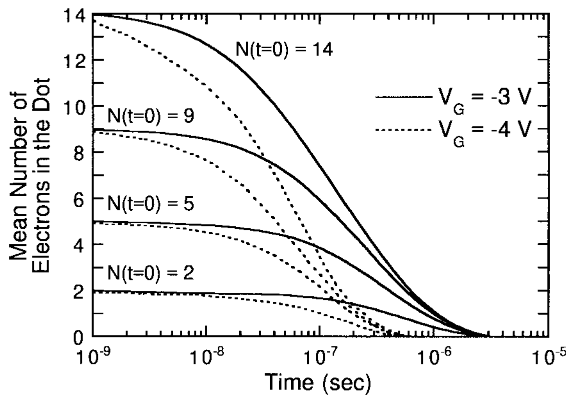


Fig. 28. Average number of electrons on a 6 × 10× 10 nm dot ($t_{inj} = 1.5$ nm) as a function of time during an erase operation for varying gate voltage and initial electron population.

what simulation indicates (on the order of ms), including the devices used in Table 9. Since volatility will be directly related to erase time, this means the retention times in the table may be longer than what would be achieved in a device with a more ideal erase time. The cause for this discrepancy is still unknown, but it is possible that the electrons are relaxing into a lower energy state in the dot, possibly even a trap state in the bandgap, after injection, resulting in longer retention and erase times.

Finally, the precise charge control afforded by the small size of the dots could be utilized to realize multilevel memory storage with inherent self-convergence. Since for robustness issues it seems clear that several electrons must be stored for each level, it will still require fairly high voltages to store multiple levels. In addition, storing more electrons on the dot will increase the electric field across the injection oxide, further decreasing the retention time. So it appears it will only be feasible with thicker oxides (>5 nm), and in systems that can trade off the higher power required for the added density obtained.

Given these performance constraints, two basic uses for the quantum-dot memory seem likely: a faster, volatile storage element (utilizing a thin tunnel oxide) with virtually unlimited write/erase capability, and a slower, nonvolatile storage element (utilizing a thick tunnel oxide) with the possibility of multilevel storage, but with limited cyclability. Note that the latter option is basically just the final scaling of a "standard" FLASH memory, while the former offers a new set of tradeoffs that would move it outside of the normal FLASH markets. The design choices would ultimately depend on the application.

In both flavors, the devices offer a highly dense memory. Since only a single-element is necessary to construct the memory cell, a $4F^2$ array layout is easily feasible in a planar technology, which would pack 5 Gbyte of storage into a square cm at F = 25 nm, and even higher densities are possible if a vertical structure is adopted (e.g., [151]). In addition, the lack of a storage capacitor makes the devices ideal for merging with logic, since the processing is so similar. This is advantageous since a high bandwidth path between the memory and the processing unit is desirable, especially as the relative difference between processor speed and memory access time continues to widen. A future processor chip would continue to require a high-speed SRAM cache to boost performance, but in addition it could contain all of the memory currently stored in separate chips, composed instead of thin-oxide, quantum-dot memories, forming a very powerful single-chip system.

IV. Applications for Nanoscale CMOS

The high integration density of CMOS, its relatively low power dissipation, and steady reductions in the cost of manufacturing have brought it to the point of being the single most important semiconductor technology in the world today. It is used in applications too numerous to mention and seems likely to continue growing in popularity. The following sections highlight a few of the possible future uses of CMOS in the nanoscale regime, first by extrapolating present logic and memory technology and then by exploring its use in analog RF circuitry and in imaging circuits.

A. Logic and Memory Applications

Digital CMOS is already widely used for general-purpose processors (GPP's), memory, and ASIC's and has made possible a wide range of interesting applications, from high-end workstations to personal digital assistants to wristwatches. The viability of many of these applications is strongly dependent on the low power dissipation and/or low cost of silicon CMOS. The low power aspects are especially important to portable applications, while the cost issues are important for large-market consumer applications.

Applications of future nanoscale CMOS are expected to be similarly dependent on power dissipation and cost issues. The power dissipation issue is addressed in Fig. 29, which shows the potential range of computation energy out to the presently perceived end of CMOS scaling. The data points represent a variety of general-purpose microprocessors and digital signal processors (DSP's) described in recent ISSCC proceedings. To facilitate comparisons among them, the energy per instruction shown on the vertical axis is taken to be the published power dissipation divided by the peak possible rate of executing instructions. For example, a four-issue superscalar processor is considered to have a peak execution rate of four times the clock frequency. The dashed lines show the expected scaling of energy per computation, based on the data in Fig. 2 and adjusted to the magnitude of these data points. Once again, both the high-performance and low-power options are shown to give some feel for the sensitivity of the energy consumption to the design goals. This plot shows a very wide range in energy consumption, from simple low-precision DSP's at the low end to high-performance general-purpose superscalar processors at the high end. Even at a fixed technology generation, the energy per instruction spans three orders of magnitude. Of this range, roughly half is attributable to power-delay tradeoffs in architecture and circuit family selection, and the other half is associated

TSMC Exhibit 1029

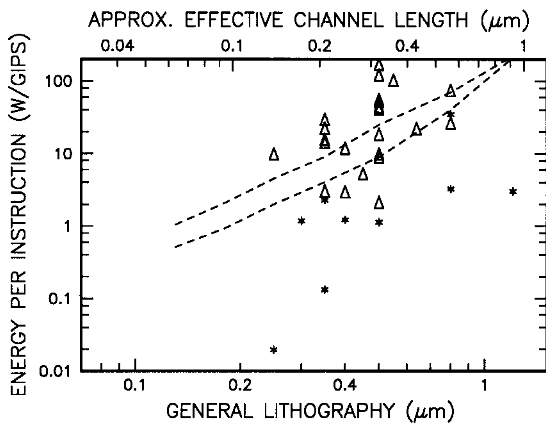


Fig. 29. Plot of energy per instruction versus general lithographic feature size. Data points are taken from recent ISSCC publications. The triangles represent general-purpose microprocessors, and the asterisks are various DSP chips. The dotted lines show the expected high-performance and low-power scaling behavior, based on the data in Fig. 2. Approximate effective channel length is indicated on the top axis, but the correspondence with general lithographic ground rules is inexact, since it varies depending on the manufacturer and the process.

with the complexity of the processor. GPP's with high data precision dissipate much more energy per operation than more specialized lower data-precision DSP chips. This plot suggests that when scaling has run its course over the next 10–15 years, the ultimate power consumption rate for GPP's will be in the vicinity of 0.1–1 W/GigaOps/s. Though some improvements on this may be possible, it does not presently appear likely that this figure can be improved by orders of magnitude using any foreseeable CMOS technology. The plot also suggests that when scaling reaches this end point, several more orders of magnitude improvement in power consumption rate can be obtained by converting applications from GPP's to special-purpose DSP-like processors wherever possible. Such processors may be able to reach 1–10 mW/GigaOps/s. Thus, even when the end of scaling is reached, circuit, system, and algorithm design ought to be able to yield significant additional progress.

Most digital CMOS applications use data that must be accessed from a memory, which can also be a significant source of power dissipation. Most of this dissipation is associated with charging and discharging long wires, which typically have capacitance of at least 0.2 pf/mm, although this could be reduced somewhat by using lower permittivity insulating material. For very large systems the size of the memory might result in a long data path and potentially high dissipation, but the use of caches reduces the average transmission distance for data, as can subdivision of the memory for parallel processing. In the case of caches, if the miss rate is not too high most of the energy is associated with the first few levels, which usually are or will be incorporated onto the chip along with the processor. In this case the memory access energy effectively scales with the processor technology, and as a first approximation we consider it to be included into the energy estimates in Fig. 29.

Some of the cost issues associated with nanoscale CMOS have already been discussed in Section II-B. For purposes of estimating the cost of future nanoscale CMOS, it suffices to note that cost depends on technology maturity, silicon area, and the value of intellectual property contained in the design. Assuming that the future mix of high and low value-added designs is similar to that at present, and projecting the (low) historical rate of cost increase out to the end of scaling, we expect silicon chips to cost \$10–100 per cm^2. The chip cost for a particular technology generation will continue to drop as manufacturing technology matures and the chips become a commodity.

Table 10 characterizes a range of possible applications for nanoscale CMOS. The numbers shown are only rough estimates for the memory and computation rates that will be needed for these applications. Data precision is a qualitative indication of the number of bits of accuracy required of the processor for most of the operations in the given application. Power requirements are very rough approximations, estimated from the required computation rate and the preceding estimates of power consumption rate at the limits of scaling. The necessary silicon real estate is estimated using 0.01 cm^2/GIPS for processing (based on scaling) and 1.4 cm^2/Gbytes for memory (based on the SIA roadmap at 2009). The following comments relate to the entries in this table.

Speech-to-text software for GPP's already exists. By the end of scaling this application can be expected to be so inexpensive and use such little power when implemented on dedicated hardware that it could be included in almost anything. Real-time language translation is a natural extension of speech recognition, probably requiring two language models and more complex algorithms.

Video communication applications are represented in the table by their most computationally intensive component—encoding. Video encoding is necessary to reduce transmission bandwidth and storage requirements, but it requires many low-precision operations, especially for motion estimation. The computational requirements given here are based on hierarchical search schemes. Four different video quality options are shown to cover the wide range of applications that are possible [152], [153].

The imagined two-way video wristwatch is an example of an application that would need to run on very low power. This particular application may or may not be such a great idea, but it exemplifies some of the issues with ultralow power interactive gadgets. For present-day wristwatches, with their year or longer battery life, the power dissipation is generally ≤ 1 μW. Even with larger higher energy-density batteries, and/or more frequent replacement, it would still appear desirable to keep the average dissipation below 10 μW. If the device were only in use 5% of the time, then an active power of roughly 200 μW could be tolerated. As indicated in the table, this power level will probably be possible, but only for the lowest resolution video compression considered.

In addition to the video codec processing, such a device is constrained by the fundamental principles of RF data trans-

TSMC Exhibit 1029

Table 10
Selected Potential Applications and Their Requirements (Power and Area are Estimates for silicon CMOS at the Limits of Scaling; Power Is Estimated for Both GPP's and for Special-Purpose DSP-Like Processors)

Application	Memory (GB)	Comp. Rate (GIPS)	Data precision	Power (W)		Si Area (cm^2)
				GPP	DSP	
speech recognition (to text)	0.01-0.1	0.1-1	low	0.1	0.001	≤ 0.1
real time language translation	0.1	1-10	low	1	0.01	0.2
video encoding						
QCIF (174 x 144, 10fps)	< 0.01	0.03-0.1	low	-	0.0002	0.01
CIF (352 x 288, 30fps)	< 0.01	0.5-5	low	-	0.005	0.03
CCIR 601 (720 x 480, 30fps)	< 0.01	2-20	low	-	0.02	0.07
Very high res.(1920 x 1200, 30fps)	0.01	20-200	low	-	0.2	0.6
2-way video wrist watch	0.01	0.05	low	-	0.0002	0.01
PDA	0.1	1-10	low	-	0.01	0.2
Tablet	0.1-1	10-100	low	-	0.1	1
factoring 512 bit numbers	1	4000	high	1000	10	40
Deep Blue chess	3	10000	low	3000	3	100
QM-based device simulation	10	100	high	30	-	15
PetaFLOPS computing challenges	3×10^4	10^6	high	10^6	-	5×10^4

mission [154], which dictate that the minimum transmitter power required to transmit data at a rate B is given by

$$P = \frac{A}{\epsilon_x} M \, \mathrm{SNR}_{\mathrm{req}} k_B T F_R B \left(2\pi \frac{df}{c}\right)^2$$

where A is a factor containing circuit losses divided by antenna gain, ϵ_x is the efficiency of the transmitter's power amplifier, M is extra signal margin to accommodate fading, $\mathrm{SNR}_{\mathrm{req}}$ is the required SNR for adequate signal detection, k_B is Boltzmann's constant, T is the temperature of the receiving antenna, F_R is the noise figure of the receiver, d is the distance between transmitter and receiver, f is the carrier frequency, and c is the speed of light. Fig. 30 shows the results of an optimistic application of this formula to the two-way video wristwatch. It shows that for carrier frequencies below 100 MHz it should be quite possible to keep the transmitter power below 1 μW and still have reasonably long-range communication. This should be quite satisfactory for use in the home or office or other places where the cell size could be small and there might be few users within each cell radius. For use on the streets of a busy city, or perhaps in a sports arena, however, this solution will not work. In these cases one must have many channels available because of the many possible simultaneous users, and this requires the use of high carrier frequencies. For example, 5000 channels might well require f =10 GHz, since most of the spectrum is unavailable due to prior allocation. This together with a 100-m cell size (which would be hard to shrink in large public locations) would require 1 mW of transmitter power, which substantially exceeds the power budget. The most likely solution would be to lower the data rate to around 10 Kb/s (since the power is proportional to the data rate) by using even more lossy video compression algorithms and transmit very low quality video in locations requiring large numbers of channels. These considerations suggest that the two-way video wristwatch is at the very limits of what may be possible, and that it, and possibly many other portable applications, is not so much limited by nanoscale CMOS technology as by physical limitations on data transmission.

Ultralow power applications can also be limited by the requirements of interacting with humans. The imagined wristwatch, for example, needs an electronic camera (see Section IV-B2), a video display (see the following paragraph), and audio output (which would probably require an earphone to keep the power down to something reasonable).

The personal digital assistant (PDA) and tablet are progressively more powerful computational devices. The PDA ought to be capable of speech recognition and perhaps language translation, too, while still maintaining a low power budget of order 10 mW. Respectably high quality video compression/decompression should also be within its domain, and RF communication should be much less constrained than for the wristwatch. The tablet is usually thought of as placing special emphasis on the display, but even with a notebook-size 300 dpi color display with perhaps 25 Mpixels, it would have enough computational power to handle video compression/decompression. Power

TSMC Exhibit 1029

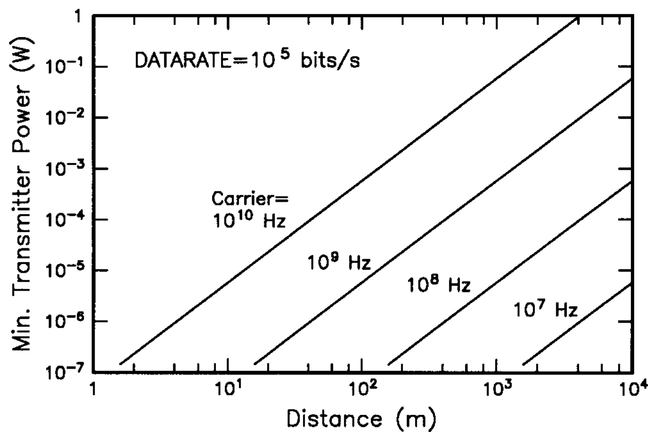


Fig. 30. Plot of minimum transmitter power versus communication distance for various carrier frequencies. The calculation optimistically assumes a data rate of 100 Kb/s, $\epsilon_x = 0.5$, $A = 1$, $M = \mathrm{SNR}_{\mathrm{req}} = 10$, $F_R = 4$, and $T = 300\mathrm{K}$. Current design practice would usually require 10–100 times more power than this to accommodate more fading margin and circuit losses.

consumption in the display drivers is not included in the table and would have to be managed very carefully in these devices, since it could easily exceed the dissipation of the processors. A rough optimistic estimate suggests that the tablet display drivers might dissipate 0.1–1 W. This would scale with the number of pixels, the linear dimension of the display and the frame rate, so that the PDA display might only require 10 mW, and the low resolution video wristwatch might get by at only 10 μW.

The factoring example is intended to be representative of cryptography applications. The requirements are estimated for factoring a 512-b integer in one day, using the number field sieve algorithm [155], and suggest that such factoring should be readily possible, though probably not on the desktop.

Deep Blue-level chess playing is an extreme example of game-playing applications. In this case it has already been demonstrated that special-purpose hardware is more effective than usual in reducing power and cost compared to implementing the same algorithms on GPP's. The most common game application will undoubtedly be video games. Three-dimensional virtual reality video games using hundreds of giga-operations per second (GOPS) ought to be possible, limited mostly by cost considerations in the consumer market.

Device simulation using a full quantum mechanical treatment will become very important over the next decade as the search for new technologies expands. This example of the use of computers in engineering suggests that it should become possible to simulate many complex engineering problems on desktop computers.

PetaFLOPS computing encompasses a wide array of applications that are already being imagined [156], even though they are not yet possible. These include structural biochemistry and the design of biomedical molecules, climate and ocean modeling to evaluate the effects of pollutants, and turbulence studies related to aeronautics, astrophysics, thermonuclear reactions, and weather. Projecting out to the limits of scaling, PetaFLOPS computers ought to become roughly equivalent to today's supercomputers in terms of price and availability. It does not appear likely that CMOS will ever be able to make such computers as common as PC's.

B. Nondata Processing Applications

Many nondata processing applications of CMOS (e.g., smart power electronics, analog/digital mixed signal communication circuits, hard disk drive read/write channels) face a unique set of problems in the era of nanoscale CMOS because the technology is developed with digital data processing in mind. Development of new generations of CMOS technology only takes into account such performance gauges as power, delay, energy, and reliability as applied to logic and memory applications. Incorporating "other applications" in the "standard" CMOS process is usually a variation on the baseline process more or less as an afterthought as economics allows. Such "other applications" that have integrated themselves into the digital world over the years will have to find a way to incorporate their device needs into future nanoscale MOS devices.

Although each application has its own set of constraints, the essence of their commonality can be extracted by examining some particular examples. In Sections IV-B1 and IV-B2, we use CMOS RF and CMOS image sensor technologies as vehicles to illustrate some of the issues involved. Both technologies would be required for the two-way video wristwatch example described in more detail in Section IV-A. Although the issues discussed are specific to RF and imaging applications, such issues often have similar counterparts in other applications.

1) Sub-100-nm CMOS for RF and Microwave: Seventy-nanometer CMOS has cutoff frequency reaching 150 GHz and inverter gate delay around 15 ps at a supply voltage of 1.5 V. SOI CMOS provides about 25% further improvement in inverter delay, which can get as low as 10 ps at V_{dd} above 2 V. The high device speed suggests that CMOS has potential for future broadband multimedia applications [157]. For example, the short gate delay makes possible a three-stage inverter voltage-controlled oscillator (VCO) operated around 10 GHz and a TSPC frequency divider above 7 GHz [157]. The f_T data in Fig. 31 suggest that even higher speed of operation is possible using analog techniques [158]. It is worth mentioning that MOSFET's with higher f_T than BJT's have long been known, but at much higher power consumption due to low g_m/I_d [159], [160]. For 100-nm MOSFET's, f_T is still as high as 40 GHz when the gate is biased near the threshold voltage, where g_m/I_d can be larger than 10 V^{-1}. Though g_m/I_d is still lower than in BJT's, the gap in performance/power tradeoff may be further narrowed by the fact that MOSFET's can be operated at lower voltages than BJT's. Deep into the subthreshold regime, where both f_T and $f_{\max}$ drop quickly, sub-100 nm MOSFET's may still provide enough bandwidth for many applications. One should note that near V_T the current–voltage relation approximately follows

TSMC Exhibit 1029

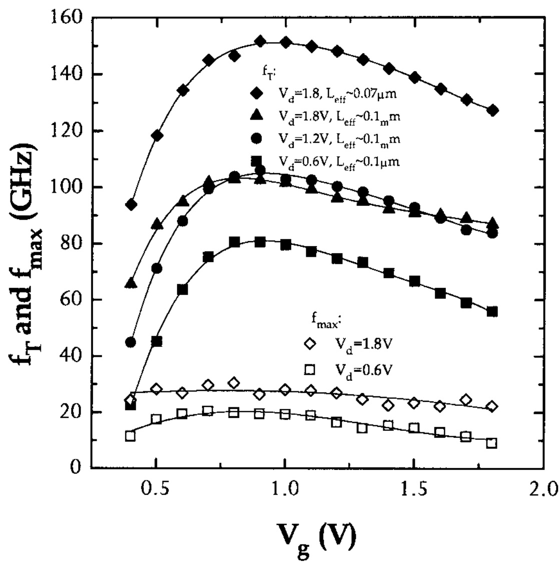


Fig. 31. f_T and $f_{\max}$ for different channel lengths and gate and drain bias conditions. After Wann *et al.* [157].

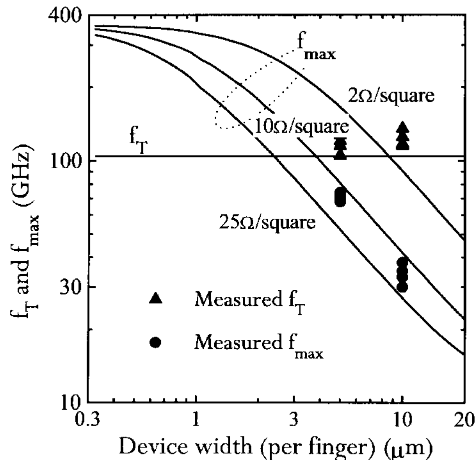


Fig. 32. Layout constraints for high $f_{\max}$ with different gate sheet resistances. After Wann *et al.* [157].

the form ln(1 + e^x) [111], which requires more careful device modeling.

Due to velocity saturation and nonscaling of parasitics in CMOS below about 150 nm, f_T is expected to scale as $1/L$, rather than $1/L^2$ as in long-channel devices. For RF applications, $f_{\max}$ is as important as f_T. With high silicide resistance, $f_{\max}$ is often gate resistance (R_g) limited, and $f_{\max}$ can be improved simply by reducing the device width per finger in the layout (Fig. 32). Currently, the gate-to-drain capacitance (C_{gd}) is playing a stronger role in affecting $f_{\max}$ than the output resistance r_o. But the trend is expected to change in further scaling since short-channel effects will become very severe. $f_{\max}$ will eventually be limited by r_o and R_s at negligibly small R_g's. In the 25-nm generation of CMOS, N-MOSFET's will likely have f_T over 300 GHz. With proper design optimization, it is possible to have $f_{\max}$ around 300 GHz, as well.

MOSFET's generally have higher $1/f$ noise than MESFET's and BJT's, due to carrier transport near the Si/SiO_2 interface. Such low-frequency noise can be reduced by improving the interface trap density or by using buried-channel devices, but for applications demanding the lowest phase noise, MOSFET's might not be the best candidate. However, for less demanding applications, MOSFET's might be good enough, particularly if the impact of phase noise can be reduced by proper design. For higher frequency noise, MOSFET's with minimized gate resistance should be comparable to, or only slightly worse than, MESFET's and BJT's at room temperature and practical drain voltages (>0.5 V).

In the nanometer regime, the square law region of long-channel MOSFET's has almost disappeared. One can estimate the third-order intermodulation interception input from device I_d–V_g characteristics. The maximum input voltage is above 2 V when biased at maximum g_m and about 1 V when biased near the threshold voltage. It is interesting to note that nanometer MOSFET's might not be able to provide the nonlinearity which some applications require, such as in certain mixers and multipliers.

From the integration point of view, substrate effects have been viewed as one drawback of silicon devices. However, progress in wafer bonding technologies and substrate micromachining might provide solutions. Another problem is that device design for digital CMOS focuses on enhancing $I_{\rm on}$ and g_m but pays little attention to r_o and noise. Also, the fixed $I_{\rm off}$ constraint for digital circuits tends to be an overkill for analog applications. The V_{dd} specifications of sub-100 nm digital CMOS leave very little head room for analog circuits. With V_{dd} at or below 1.5 V, the voltage drop in the tail current source, cascode, or source follower can be significant. Thus, higher operating voltage is usually desirable in analog functions and can be acceptable since devices are often stacked. Eventually a different device design for RF might be needed.

2) Imaging: For image sensors based on CMOS [161]–[165], the pixel is composed of a photosensitive element and in-pixel active MOSFET's to perform the four vital functions of: 1) photon to electron conversion; 2) pixel selection; 3) photosensitive element reset; and 4) photocharge to voltage (or current) amplification. The photosensitive element is typically a depletion region (formed either by a pn junction or electric field-induced by an MOS gate) that collects electrons generated by photons absorbed in the silicon.

Image sensors belong to a category of devices where the physical device size cannot be arbitrarily reduced. The resolution of the optical system that focuses the image onto the image plane determines the lower bound of the pixel size. The Raleigh criterion determines the minimum resolvable distance (δ) of two points limited by diffraction ($\delta = 1.22\lambda \times f$-number, where λ is the wavelength of the light). With a typical f-number of $f/8$ and 550 nm light (green), the minimum resolvable distance is 5.4 μm. Therefore, the pixel size will be limited to about 5 μm if the sole driving force for pixel size reduction is improved resolution. Considerations such as chip size and chip cost may drive the pixel toward smaller sizes [166], [167].

TSMC Exhibit 1029

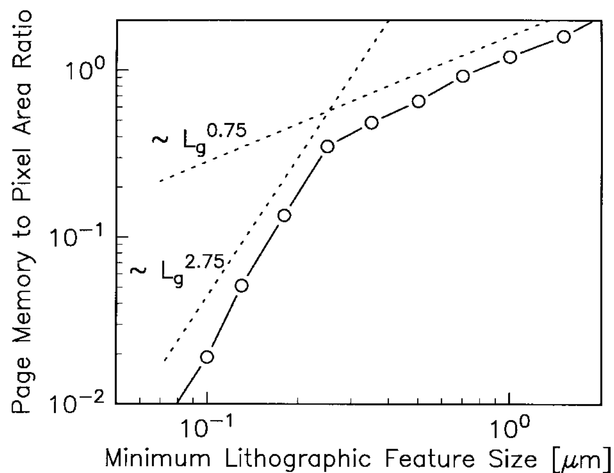


Fig. 33. The ratio of the chip area that would be occupied by a full page of 8-b DRAM to the area occupied by the pixels. This ratio is independent of the number of pixels on the sensor as a function of the minimum feature size. The change of slope occurs at the point where the pixel size does not decrease with lithographic feature size due to limits in the optical system. After Wong [164].

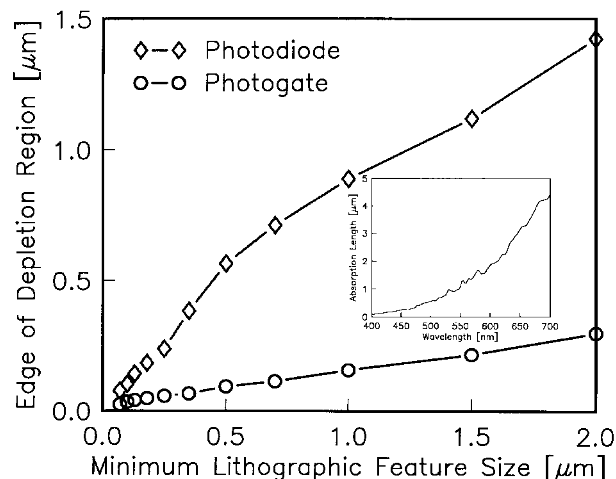


Fig. 34. The location of the edge of the depletion region as a function of the minimum lithographic feature size, L_g. After Wong [164].

The benefits of device scaling on CMOS imagers are higher levels of integration and lower power consumption. On the pixel level, more sophisticated amplification circuit or even per-pixel signal processing operation (e.g., per-pixel analog-to-digital conversion [168]) can be envisioned. On the system level, one can incorporate analog-to-digital conversion, DSP functions (e.g., color processing, image segmentation, image compression) [169], [170] or memory. Because the size of the pixels cannot be arbitrarily reduced much beyond the optical limits, the chip area cost of including such on-chip functions becomes minimal as CMOS scales into the nanometer regime. This is illustrated in Fig. 33, where the ratio of the chip area that would be occupied by a full page (corresponding to the number of pixels in the imaging array) of 8-b DRAM to the area occupied by the pixels is plotted versus lithographic feature sizes. In the nanometer regime, it only takes a tiny chip area to incorporate on-chip system functions.

As CMOS scales into the nanometer regime, one needs to examine whether CMOS can continue to perform the four functions required of a pixel given the change of device structure dictated by scaling. This is discussed in more detail in [164] and the salient features are summarized here.

Device scaling dictates that depletion regions be scaled linearly with the gate length. Fig. 34 shows the edge of the depletion region (below the silicon surface) as a function of the lithographic feature size. Comparing Fig. 34 with the absorption length of visible light (see inset of Fig. 34), it is apparent that for CMOS imagers, most photocarriers are not generated in the depletion region, but rather deeper in the silicon bulk and subsequently collected via carrier diffusion. The minority carrier diffusion length decreases almost linearly with the lithographic feature size because the minority carrier lifetime is inversely proportional to the substrate doping at the doping levels of interest, while the mobility decreases relatively slowly with increasing substrate doping. Thus, photo-generated carriers will increasingly be lost to recombination (decreased quantum efficiency) because they are generated deep in the silicon compared to the shrinking depletion regions and diffusion lengths.

Many nanoscale MOS devices call for using very thin (less than 10 nm) layers of silicon. This poses a very serious problem for devices which depend on having a large region of silicon for their operation, for example, image sensors and power devices. At 10-nm thickness, the active silicon will be mostly transparent to visible light. Photosensitive elements that have inherent gain mechanisms need to be devised in lieu of conventional photodiodes or photogates where there are no photon-to-charge conversion gain. Another alternative is to devise a photon collection region that is not tied to the thickness of the MOSFET channel region.

Circuits that convert photons to voltages (or currents) are essentially analog circuits. They suffer from the same limitations imposed by a lower power supply voltage (V_{DD}) in the nanoscale CMOS regime as discussed at the end of Section IV-B1. Eventually, it is possible that different types of pixel circuit or modified process technologies will be required to enable imaging functions on CMOS chips.

V. Discussion

This paper began by first attempting to address the questions outlined in the introduction, namely, what are the apparent limits of CMOS technology, how will CMOS be extended into the nanometer regime, and what sort of applications is it expected that CMOS will and will not be able to handle? The viewpoints presented in this paper are necessarily biased by our own experiences and the environment in which we find ourselves (namely, a large, industrial research laboratory). Nevertheless, the discourse in this paper, when taken in its proper context, should shed light on the place of CMOS in the next 15–20 years.

The history of CMOS and its successful scaling over the past 25 years as reviewed in Section II suggest the re-

TSMC Exhibit 1029

markably good fortune that the same basic device structure, after some serious engineering, has provided generation after generation of density/power/performance improvements. As the nanometer regime is being penetrated more deeply, it is apparent that electrostatic control of the channel and use of dopants to achieve precise control of the field configuration are reaching their limits. Silicon dioxide, the key to the past success of CMOS, is finally reaching its limit, too, and failing a major breakthrough in alternative gate dielectrics will cause scaling of bulk CMOS to end at a gate length of about 50 nm. Already, in research labs around the world, the last generation of bulk CMOS is being explored. Device structural changes (to thin SOI, ground plane FET, and DG FET) will alleviate some constraints and perhaps extend CMOS for one to three more generations but carry with them a different set of difficulties. Such difficulties include controlling the silicon channel thickness and fabricating three-dimensional-type devices using essentially planar, layer-by-layer-type processes. Mask layout for some of these new structures would most likely be incompatible with existing designs, necessitating large redesign expenses. Silicon technology which has used basically the same, limited, material set through all these years will probably require new materials, especially for the gate, to implement these new structures. Once scaling stops, materials such as SiGe might still be used to increase performance.

The analyses of dopant fluctuation effects in Section II-D3 suggest that the industry projections for the reduction of supply voltages are too optimistic considering the expected parameter uncertainty due to fundamental physics, such as dopant fluctuation and process tolerances coupled with the vast number of devices per chip. This difficulty arises in part from the expectation that the power supply voltage will be reduced to a value close to the threshold voltage. Solutions might be found either by keeping a reasonable value of $V_{DD} - V_T$ or alleviating dopant fluctuation effects through innovative device structures such as those described in Section III-A.

Quantum effects, which loom so large in other nanotechnologies, are for the most part undesirable in CMOS nanotechnology, especially logic technology, since they limit the minimum gate oxide thickness, silicon thickness (for SOI), and depletion layer thicknesses. Single electron effects, especially for memory, could prove useful in standardizing the charge state of the floating-gate-type memory. There seems to be little prospect of integrating quantum-effect devices, such as lasers, into CMOS, but this does not preclude detector arrays (see Section IV-B2), and the possible hybrid attachment of III–V devices onto a CMOS chip.

Three-dimensional integration, when implemented by stacking separate planes of active devices interconnected by multiple wiring planes, does not seem to offer a significant advantage because of the blockage of wiring planes. More innovative means of meshing the active devices and their wiring in three dimensions seem necessary, although no promising concept exists today. The merging of memory and logic devices (as well as other analog and sensing functions) on the same chip appears to be a logical extension of technology integration, providing the benefits of flexible system partitioning. The novel memory devices described in Section III-B1, for example, can be fabricated almost identically to a logic transistor. While the technical viability of merging different technologies into a single, integrated technology is not in doubt (at least in principle), whether this technology integration will happen will mostly be dictated by economics and market demand. It is expected that bulk CMOS in production will continue to improve at its historical pace in the coming eight to ten years. As the limitations of bulk CMOS become apparent, devices based on the more innovative structures described in Section III-A will begin to be used. Progress may then slow down (compared with the present projections [6]) due to the difficulty of bringing these new structures and materials into manufacturing.

Section IV examined the possible applications of CMOS in the nanometer regime. Power consumption dominates most of the considerations of what is achievable and what is not possible with the technology. Perhaps the most important observation from past experiences is the ability to achieve low power consumption (and hence extend the applicability of CMOS to various applications) by custom low-power design of special processors. This will perhaps herald the end of the era of GPP's (such as the x86 microprocessors) for many applications in much the same way that the personal computer erased the dominance of the mainframe computer. In their place will be application-specific microprocessors (ASμPs). This is especially true for future compute-enabled daily appliances and personal gadgets where a small form-factor and mobile attributes call for high levels of integration and the lowest power consumption. Nanoscale CMOS opens up new possibilities such as RF applications, while presenting new challenges to existing applications such as imaging. Our analyses show that a two-way video wristwatch (which has always been the fancy of many science/engineering enthusiasts) appears possible, albeit with limited video quality. This, and possibly many other portable applications, appears to be at the limit of what may be feasible. PetaFLOPS-scale computing ought to be within the reach of nanoscale CMOS supercomputers, while high-end workstations should be capable of full quantum mechanical simulation of semiconductor devices.

What applications will nanoscale CMOS not be capable of handling? The discussion of applications has not directly answered this question. The problem is that there is a large set of applications for which it is difficult to arrive at even slightly accurate estimates of computing resource requirements. These applications include things like chauffeuring (driving a car), piloting fighter jets in combat, robotic maids and butlers, and electronic factory workers or, perhaps, IC designers. To be fully effective, these "devices" ("creatures?") need nearly human-level intelligence, at least within their individual specializations. Much work is still needed on the algorithms and programming methods

TSMC Exhibit 1029

necessary to create such intelligence, but as an upper bound we can consider emulating the human brain itself. (Note that this may be analogous to requiring flapping bird-like wings for flight.) The computational requirements of the brain are thought to be around 10^{17} low precision operations per second and perhaps 10^{14} bytes of memory [171]. According to Section IV-A, this would require 10^6 cm^2 of silicon at an expense of 10 M just for the hardware in a mature technology. The power dissipation would be at least 100 KW. We conclude that unless more efficient algorithms and/or circuits are found, nanoscale CMOS will be unable to bring these high-intelligence applications to the consumer market. They will only be of use to the high-end military and industrial sectors. In all probability this is a very pessimistic assessment. Clever techniques will probably be found to reduce substantially the computational requirements for these applications, but it remains to be seen what will be possible. On the other hand, it is interesting to note that the energy per arithmetic operation at the limits of scaling (1 mW/GigaOps/s) is equivalent to 10^9 kT at room temperature, which is well above physical limits on dissipative computation. This suggests that there is still room for new technologies that can lower the energy consumption of computation.

VI. Conclusions

CMOS technology is approaching a scaling limit dictated not so much by the ability to make a small FET that switches fast, but by the ability to make such an FET that also has extremely small leakage current when turned off, that has a threshold voltage that is not too dependent on gate length, that does not suffer too much from parameter-induced fluctuations, and that can be interconnected three dimensionally with multiple levels of wiring. The industrial base and infrastructure of CMOS technologies are enormous compared to other emerging, competing technologies. As CMOS progresses on a three-year (or less) cycle of 2× increase in performance and 4× increase in the number of devices per chip, it will be difficult for alternative technologies to leapfrog the incumbent technology. It is seldom the case that completely new technologies replace existing technologies serving exactly the same functions. Successful emerging technologies must provide new applications desired by a large market which existing technologies cannot provide. For example, the transistor provides improvements in speed, power, size, and reliability that vacuum tubes cannot achieve. Viewed in this context, it is clear that CMOS has set a very high standard against which contending nanotechnologies must be compared. Nevertheless, as indicated in Section V, there are many applications desired by society that may not be fulfilled by conventional CMOS technologies alone, even in the nanoscale regime. It is those applications on which aspiring technologies should focus attention.

CMOS technology has had 30 years of continuous, exponential development. Today, we are beginning to see the end of this road, for device technologies, in perhaps five to six generations and about 15 years into the future (according to the 2× performance improvement every three years projection) or maybe even longer. While this may seem very far off, the lead time between research and ubiquitous deployment in the industry is such that research is already being carried out today, in universities and industrial laboratories, for prototypes of this final generation. Yet, we cannot currently see another nanotechnology to continue where CMOS leaves off—the requirements for future generations of computing technologies are so demanding. Without the driving force of continued exponential growth, a large infrastructure of education, research, and development may be left stranded without much economic underpinning. This worst-case scenario should be of concern to educators and technology leaders. On the other hand, concomitant with the advances in device technologies are advances in circuit and system design, communication technologies, and data-storage technologies which will continue to provide generations of performance/efficiency enhancements even after device technologies cease to improve at their historically phenomenal pace. In fact, as illustrated in Section IV-A, there is plenty of room for improvement in circuit and system design for a given device technology. To paraphrase Feynman's famous quote ("There is plenty of room at the bottom"), it appears that "there is plenty of performance to be milked from nanoscale CMOS."

Acknowledgment

The authors would like to acknowledge the contributions of the Silicon Innovation Laboratory for fabricating the devices described in this paper. The authors benefited from discussions with R. Dennard, T. Ning, and Y. Taur. Critical reading of the manuscript by R. Dennard is much appreciated. J. J. Welser would like to acknowledge S. Tiwari and F. Rana for their large contributions to the technical aspects of this paper. H.-S. P. Wong would like to acknowledge K. Chan and Y. Taur for collaboration on device fabrication. J. J. Welser and H.-S. P. Wong would like to thank J. Benedict for TEM analyses. D. J. Frank would like to acknowledge useful discussions with B. Gaucher, C. Gonzalez, P. Bose, and C.-K. Hu, and valuable assistance in locating economic data from M. Cowan and M. Moser. The management support of J. Warlaumont is greatly appreciated.

References

[1] B. Davari, R. H. Dennard, and G. G. Shahidi, "CMOS scaling, the next ten years," *Proc. IEEE,* vol. 83, p. 595, 1995.

[2] Y. Taur, D. Buchanan, W. Chen, D. Frank, K. Ismail, S.-H. Lo, G. Sai-Halasz, R. Viswanathan, H.-J. C. Wann, S. Wind, and H.-S. Wong, "CMOS scaling into the nanometer regime," *Proc. IEEE,* vol. 85, p. 486, Apr. 1997.

[3] S. Asai and Y. Wada, "Technology challenges for integration near and below 0.1 μm," *Proc. IEEE,* vol. 85, p. 505, Apr. 1997.

[4] R. Dennard *et al.,* "Design of ion-implanted MOSFET's with very small physical dimensions," *IEEE J. Solid-State Circuits,* vol. SC-9, p. 256, 1974.

[5] D. J. Frank, "Application and technology forecast," in *Low Power Design in Deep Submicron Electronics,* vol. 337, W.

TSMC Exhibit 1029

Nebel and J. Mermet, Eds. Dordrecht, The Netherlands: Kluwer, 1997, pp. 9–44.
[6] Semiconductor Industry Association (SIA), *The National Technology Roadmap for Semiconductors.* [Online.] Available WWW: http://www.sematech.org/public/roadmap/index.htm.
[7] G. E. Moore, "Lithography and the future of Moore's law," *Proc. SPIE*, vol. 2437, pp. 2–17, 1995.
[8] B. Davari, C. Koburger, T. Furukawa, Y. Taur, W. Noble, A. Megdanis, J. Warnock, and J. Mauer, "A variable-size shallow trench isolation (STI) technology with diffused sidewall doping for submicron CMOS," in *Proc. Int. Electron Devices Meeting,* 1988, p. 92.
[9] W.-H. Chang, B. Davari, M. R. Wordeman, Y. Taur, C. C.-H. Hsu, and M. D. Rodriguez, "A high-performance 0.25-μm CMOS technology—I: Design and characterization," *IEEE Trans. Electron Devices,* vol. 39, p. 959, Apr. 1992.
[10] B. Davari, W.-H. Chang, K. E. Petrillo, C. Y. Wong, D. Moy, Y. Taur, M. R. Wordeman, J. Y. C. Sun, and C. C.-H. Hsu, "A high-performance 0.25-μm CMOS technology—II: Technology," *IEEE Trans. Electron Devices,* vol. 39, p. 967, Apr. 1992.
[11] G. Sai-Halasz *et al.,* "Experimental technology and characterization of self-aligned 0.1 μm, gate-length low-temperature operation NMOS devices," in *Proc. Int. Electron Devices Meeting,* 1987, p. 397.
[12] Y. Taur *et al.,* "High transconductance 0.1 μm pMOSFET," in *Proc. Int. Electron Devices Meeting,* 1992, pp. 901–904.
[13] Y. Taur *et al.,* "High performance 0.1 μm, CMOS devices with 1.5 V power supply," in *Proc. Int. Electron Devices Meeting,* 1993, pp. 127–130.
[14] S. Ogura, P. Tsang, W. Walker, D. Critchlow, and J. Shepard, "Elimination of hot electron gate current by lightly doped drain-source structure," in *Proc. Int. Electron Devices Meeting,* 1981, p. 651.
[15] G. A. Sai-Halasz, M. R. Wordeman, D. P. Kern, S. Rishton, and E. Ganin, "High transconductance and velocity overshoot in NMOS devices at the 0.1 μm-gate-length level," *IEEE Electron Device Lett.,* vol. EDL-9, p. 464, 1988.
[16] R. Yan *et al.,* "High-performance 0.1 μm room temperature Si MOSFET's," in *Proc. Symp. VLSI Technology,* 1992, p. 86.
[17] Y. Mii *et al.,* "High performance 0.1 μm nMOSFET's with 10 ps/stage delay (85 K) at 1.5 V power supply," in *Proc. Symp. VLSI Technology,* 1993, pp. 91–92.
[18] M. Ono, M. Saito, T. Yoshitomi, C. Fiegna, T. Ohguro, and H. Iwai, "Sub-50 nm gate length n-MOSFET's with 10 nm phosphorus source and drain junctions," in *Proc. Int. Electron Devices Meeting,* 1993, p. 119.
[19] Y. Mii *et al.,* "An ultra-low power 0.1 μm CMOS," in *Proc. Symp. VLSI Technology,* 1994, pp. 9–10.
[20] H. Momose, M. Ono, T. Yoshitomi, T. Ohguro, S. Nakamura, M. Saito, and H. Iwai, "Tunneling gate oxide approach to ultra-high current drive in small-geometry MOSFET's," in *Proc. Int. Electron Devices Meeting,* 1994, p. 593.
[21] L. Su *et al.,* "A high performance 0.08 μm CMOS," in *Proc. Symp. VLSI Technology,* 1996, p. 12.
[22] C. Wann *et al.,* "High-performance 0.07 μm CMOS with 9.5 ps gate delay and 150 GHz f_T," *IEEE Electron Device Lett.,* vol. 18, p. 625, Dec. 1997.
[23] F. Assaderaghi *et al.,* "A 7/9/5.5 psec room/low temperature SOI CMOS," in *Proc. Int. Electron Devices Meeting,* 1997, p. 415.
[24] G. Timp *et al.,* "Low leakage, ultra-thin, gate oxides for extremely high performance sub-100 nm nMOS-FET's," in *Proc. Int. Electron Devices Meeting,* 1997, p. 930.
[25] M. T. Bohr, "Interconnect scaling—The real limiter to high performance ULSI," in *Proc. Int. Electron Devices Meeting,* 1995, p. 241.
[26] S. Sun, "Process technologies for advanced metallization and interconnect systems," in *Proc. Int. Electron Devices Meeting,* 1997, p. 765.
[27] C. K. Hu, D. C. Edelstein, C. Uzoh, and T. Sullivan, "Comparison of electromigration in submicron Al(Cu) and Cu thin film lines," in *Stress Induced Phenomena in Metallization, AIP Conf. Proc. no. 373,* 1996, pp. 153–68.
[28] D. Edelstein *et al.,* "Full copper wiring in a sub-0.25 μm CMOS ULSI technology," in *Proc. Int. Electron Devices Meeting,* 1997, p. 773.
[29] Zielinski *et al.,* "Damascene integration of copper and ultralow-k xerogel for high performance interconnects," in *Proc. Int. Electron Devices Meeting,* 1997, p. 936.
[30] M. Matsuura, I. Tottori, K. Goto, K. Maekawa, and M. Hirayama, "A highly reliable self-planarizing low-k intermetal dielectric for sub-quarter micron interconnects," in *Proc. Int. Electron Devices Meeting,* 1997, p. 785.
[31] G. Sai-Halasz, "Performance trends in high-end processors," *Proc. IEEE,* vol. 83, p. 20, Jan. 1995.
[32] R. Dennard, "Field effect transistor memory," U.S. Patent 3387286, July 14, 1968.
[33] T. Brunner, "Pushing the limits of lithography for IC production," in *Proc. Int. Electron Devices Meeting,* 1997, p. 9.
[34] S.-H. Lo, D. Buchanan, and Y. Taur, "Modeling and characterization of n^+- and p^+-polysilicon-gated ultra thin oxides (21–26Å)," in *Proc. Symp. VLSI Technology,* 1997, pp. 149–150.
[35] S.-H. Lo, D. Buchanan, Y. Taur, and W. Wang, "Quantum-mechanical modeling of electron tunneling current from the inversion layer of ultra-thin-oxide MOSFET's," *IEEE Electron Device Lett.,* vol. 18, p. 209, May 1997.
[36] F. Rana, S. Tiwari, and D. Buchanan, "Self-consistent modeling of accumulation layers and tunneling currents through very thin oxides," *Appl. Phys. Lett.,* vol. 69, no. 8, p. 1104, 1996.
[37] S. Tiwari, J. Welser, D. DiMaria, and F. Rana, "Currents, surface potentials, and defect generation in 1.2–1.5 nm oxide MOSFET's," in *Proc. Device Research Conf.,* 1998, p. 12.
[38] D. Frank, Y. Taur, and H.-S. P. Wong, "Generalized scale length for two-dimensional effects in MOSFET's," *IEEE Electron Device Lett.,* vol. 19, p. 385, Oct. 1998.
[39] D. Frank, S. Laux, and M. Fischetti, "Monte Carlo simulation of a 30 nm dual-gate MOSFET: How far can Si go?," in *Proc. Int. Electron Devices Meeting,* 1992, p. 553.
[40] C. Fiegna, H. Iwai, T. Wada, T. Saito, E. Sangiorgi, and B. Ricco, "A new scaling methodology for the 0.1–0.025 μm MOSFET," in *Proc. Symp. VLSI Technology,* 1992, p. 33.
[41] A. Sharma, *Semiconductor Memories: Technology, Testing, and Reliability.* Piscataway, NJ: IEEE Press, 1997.
[42] E. Adler, J. DeBrosse, S. Geissler, S. Holmes, M. Jaffe, J. Johnson, C. W. Koburger, III, J. Lasky, B. Lloyd, G. Miles, J. Nakos, W. P. Noble, Jr., S. Voldman, M. Armacost, and R. Ferguson, "The evolution of IBM CMOS DRAM technology," *IBM J. Res. Develop.,* vol. 39, nos. 1/2, p. 167, 1995.
[43] G. Bronner, "DRAM technology trends for 256 Mb and beyond," in *Proc. Int. Electron Devices and Materials Symp.,* 1996, p. 75.
[44] H. Sunami, "Cell structures for future DRAM's," in *Proc. Int. Electron Devices Meeting,* 1985, p. 694.
[45] K. Itoh, Y. Nakagome, S. Kimura, and T. Watanabe, "Limitations and challenges of multi-gigabit DRAM circuits," in *Proc. Symp. VLSI Circuits,* 1996, p. 2.
[46] P. Pavan, R. Bez, P. Olivo, and E. Zanoni, "Flash memory cells—An overview," *Proc. IEEE,* vol. 85, p. 1248, Aug. 1997.
[47] B. Hoeneisen and C. A. Mead, "Fundamental limitations in microelectronics—I: MOS technology," *Solid State Electron.,* vol. 15, p. 819, 1972.
[48] R. W. Keyes, "Effect of randomness in the distribution of impurity ions on FET thresholds in integrated electronics," *IEEE J. Solid-State Circuits,* vol. SC-10, p. 245, 1975.
[49] H.-S. Wong and Y. Taur, "Three-dimensional 'atomistic' simulation of discrete microscopic random dopant distributions effects in sub-0.1 μm MOSFET's," in *Proc. Int. Electron Devices Meeting,* 1993, pp. 705–708.
[50] T. Mizuno, J. Okamura, and A. Toriumi, "Experimental study of threshold voltage fluctuations using an 8k MOSFET's array," in *Proc. VLSI Symp.,* 1993, p. 41.
[51] V. De, X. Tang, and J. Meindl, "Random MOSFET parameter fluctuation limits to gigascale integration (GSI)," in *Proc. VLSI Symp.,* 1996, p. 198.

TSMC Exhibit 1029

[52] D. Burnett and S.-W. Sun, "Statistical threshold-voltage variation and its impact on supply-voltage scaling," *Proc. SPIE,* vol. 2636, p. 83, 1995.
[53] H.-S. P. Wong, Y. Taur, and D. Frank, "Discrete random dopant distribution effects in nanometer-scale MOSFET's," *Microelectronic Reliability,* vol. 38, no. 9, pp. 1447–1456, 1998.
[54] A. P. Chandrakasan and R. W. Brodersen, "Minimizing power consumption in digital CMOS circuits," *Proc. IEEE,* vol. 83, pp. 498–523, 1995.
[55] D. Singh, J. M. Rabaey, M. Pedram, S. R. F. Catthoor, N. Sehgal, and T. J. Mozdzen, "Power conscious CAD tools and methodologies: A perspective," *Proc. IEEE,* vol. 83, pp. 570–593, 1995.
[56] C. Piguet, "Circuit and logic level design," in *Low Power Design in Deep Submicron Electronics,* vol. 337, W. Nebel and J. Mermet, Eds. Dordrecht, The Netherlands: Kluwer, 1997, pp. 105–134.
[57] C. Svensson and D. Liu, "Low power circuit techniques," in *Low Power Design Methodologies,* P. M. Rabaey and M. Pedram, Eds. Dordrecht, The Netherlands: Kluwer, 1996.
[58] T. H. Meng, B. M. Gordon, E. K. Tsern, and A. C. Hung, "Portable video-on-demand in wireless communication," *Proc. IEEE,* vol. 83, pp. 359–380, Apr. 1995.
[59] Z.-H. Liu, C. Hu, J.-H. Huang, T.-Y. Chan, M.-C. Jeng, P. Ko, and Y. Cheng, "Threshold voltage model for deep-submicrometer MOSFET's," *IEEE Trans. Electron Devices,* vol. 40, p. 86, 1993.
[60] H.-S. Chen, C. Teng, J. Zhao, L. Moberly, and R. Lahri, "Analog characteristics of drain engineered submicron MOSFET's for mixed-signal applications," *Solid State Electron.,* p. 1857, 1995.
[61] D. J. Frank, P. Solomon, S. Reynolds, and J. Shin, "Supply and threshold voltage optimization for low power design," in *Proc. 1997 Int. Symp. Low Power Electronics and Design,* 1997, pp. 317–322.
[62] J. Burr and A. Peterson, "Ultra low power CMOS technology," in *NASA VLSI Design Symp.,* 1991, pp. 4.2.1–13.
[63] A. J. Bhavnagarwala, V. K. De, B. Austin, and J. D. Meindl, "Circuit techniques for low power CMOS GSI," in *1996 ISLPED Dig. Tech. Papers,* pp. 193–196.
[64] Z. Chen, J. Shott, J. Burr, and J. D. Plummer, "CMOS technology scaling for low voltage low power applications," in *1994 SLPE Dig. Tech. Papers,* pp. 56–57.
[65] S.-W. Sun and P. G. Y. Tsui, "Limitation of CMOS supply-voltage scaling by MOSFET threshold-voltage variation," *IEEE J. Solid-State Circuits,* vol. 30, pp. 947–949, 1995.
[66] J. Meindl *et al., Proc. Int. Solid State Circuits Conf.,* 1993, p. 124.
[67] J. Meindl, "Low power microelectronics—Retrospect and prospect," *Proc. IEEE,* vol. 83, p. 619, Apr. 1995.
[68] J. Meindl, V. De, D. Wills, J. Eble, X. Tang, J. Davis, B. Austin, and A. Bhavnagarwala, "The impact of stochastic dopant and interconnect distributions on gigascale integration," in *Proc. Int. Solid State Circuits Conf.,* 1997, p. 232.
[69] C. Mead, "Scaling of MOS technology to submicrometer feature sizes," *J. VLSI Signal Processing,* pp. 9–25, 1994.
[70] Y. Taur and E. Nowak, "CMOS devices below 0.1 μm: How high will performance go?," in *Proc. Int. Electron Devices Meeting,* 1997, p. 215.
[71] A. Auberton-Hervé, "SOI: Materials to systems," in *Proc. Int. Electron Devices Meeting,* 1996, p. 3.
[72] G. Shahidi, B. Davari, Y. Taur, J. Warnock, M. Wordeman, P. McFarland, S. Mader, M. Rodriguez, R. Assenza, G. Bronner, B. Ginsberg, T. Lii, M. Polcari, and T. Ning, "Fabrication of CMOS on ultrathin SOI obtained by epitaxial lateral overgrowth and chemical-mechanical polishing," in *Proc. Int. Electron Devices Meeting,* 1990, p. 587.
[73] A. Ogura and Y. Fujimoto, "Novel technique for Si epitaxial later overgrowth: Tunnel epitaxy," *Appl. Phys. Lett.,* vol. 56, p. 2205, 1989.
[74] ——, "Extremely thin and defect-free Si-on-insulator fabrication by tunnel epitaxy," *Appl. Phys. Lett.,* vol. 57, no. 26, p. 2806, 1990.
[75] A. Ogura, A. Furuya, and R. Koh, "50-nm-thick silicon-on-insulator fabrication by advanced epitaxial lateral overgrowth: Tunnel epitaxy," *J. Electrochemical Soc.,* vol. 140, no. 4, p. 1125, 1993.
[76] P. Schubert and G. W. Neudeck, "Confined lateral selective epitaxial growth of silicon for device fabrication," *IEEE Electron Device Lett.,* vol. 10, p. 181, 1990.
[77] S. Venkatesan, C. Subramanian, G. Neudeck, and J. Denton, "Thin-film silicon-on-insulator (SOI) device applications of selective epitaxial growth," in *Proc. Int. SOI Conf.,* Palm Springs, CA, 1993, p. 76.
[78] J. Siekkinen, G. Neudeck, J. Glenn, and S. Venkatesan, "A novel high-speed silicon bipolar transistor utilizing SEG and CLSEG," *IEEE Trans. Electron Devices,* p. 862, 1994.
[79] C. K. Subramanian and G. Neudeck, "Large area silicon on insulator by double-merged epitaxial lateral overgrowth," *J. Vac. Sci. Technol. B,* p. 643, 1992.
[80] H.-S. Wong, K. Chan, Y. Lee, P. Roper, and Y. Taur, "Ultra-thin, highly uniform thin film SOI MOSFET with low series resistance fabricated using pattern-constrained epitaxy (PACE)," in *Proc. Symp. VLSI Technology,* 1996, p. 94.
[81] G. Shahidi, C. Anderson, B. Chappel, T. Chappel, J. Comfort, B. Davari, R. Dennard, R. Franch, P. McFarland, J. Neely, T. Ning, M. Polcari, and J. Warnock, "A room temperature 0.1 μm CMOS on SOI," *IEEE Trans. Electron Devices,* vol. 12, p. 2405, 1994.
[82] J. Sleight and K. Mistry, "A compact body contact technology for SOI transistors," in *Proc. Int. Electron Devices Meeting,* 1997, p. 419.
[83] W. Chen, Y. Taur, D. Sadana, K. Jenkins, J. Sun, and S. Cohen, "Suppression of the SOI floating-body effects by linked-body device structure," in *Proc. Symp. VLSI Technology,* 1996, p. 92.
[84] J. Gautier, M. Pelella, and J. G. Fossum, "SOI floating-body, device and circuit issues," in *Proc. Int. Electron Devices Meeting,* 1997, p. 407.
[85] A. Wei and D. Antoniadis, "Design methodology for minimizing hysteretic v_t variation in partially-depleted SOI CMOS," in *Proc. Int. Electron Devices Meeting,* 1997, p. 411.
[86] T. Ohno, M. Takahashi, A. Ohtaka, Y. Sakakibara, and T. Tsuchiya, "Suppression of the parasitic bipolar effect in ultra-thin-film nMOSFET's/SIMOX by Ar ion implantation into source/drain regions," in *Proc. Int. Electron Devices Meeting,* 1995, p. 627.
[87] M. Terauchi, M. Yoshimi, A. Murakoshi, and Y. Ushiku, "Suppression of the floating-body effects in SOI MOSFETS by bandgap engineering," in *Proc. Symp. VLSI Technology,* 1995, p. 35.
[88] D. Schepis, F. Assaderaghi, D. Yee, W. Rausch, R. Bolam, A. Ajmera, E. Leobandung, S. Kulkarni, R. Flaker, D. Sadana, H. Hovel, T. Kebede, C. Schiller, S. Wu, L. Wagner, M. Saccamango, S. Ratanaphanyarat, J. Kuang, M. Hsieh, K. Tallman, R. Martino, D. Fitzpatrick, D. Badami, M. Hakey, S. F. Chu, B. Davari, and G. Shahidi, "A 0.25 μm CMOS SOI technology and its application to 4 Mb SRAM," in *Proc. Int. Electron Devices Meeting,* 1997, p. 587.
[89] Y.-H. Koh, M.-R. Oh, J.-W. Lee, W.-C. Lee, C.-K. Park, J.-B. Park, Y.-C. Heo, H.-M. Rho, B.-C. Lee, M.-J. Chung, M. H. H.-S. Kim, K.-S. Choi, K.-H. A. W.-C. Lee, K.-W. Park, J.-Y. Yang, H.-K. Kim, D.-H. Lee, and L.-S. Hwang, "l giga bit SOI DRAM with fully bulk compatible process and body-contacted SOI MOSFET structure," in *Proc. Int. Electron Devices Meeting,* 1997, p. 579.
[90] K. Mistry, G. Grula, J. Sleight, L. Blair, R. Stephany, R. Flatley, and P. Skerry, "A 2.0 V, 0.35 μm partially depleted SOI-CMOS technology," in *Proc. Int. Electron Devices Meeting,* 1997, p. 583.
[91] R. Chau, R. Arghavani, M. Alavi, D. Douglas, J. Greason, R. Green, S. Tyagi, J. Xu, P. Packan, S. Yu, and C. Liang, "Scalability of partially depleted SOI technology for sub-0.25 μm logic applications," in *Proc. Int. Electron Devices Meeting,* 1997, p. 591.
[92] F. Assaderaghi, W. Ruasch, A. Ajmera, E. Leobandung, D. Schepis, L. Wagner, H.-J. Wann, R. Bolam, D. Yee, B. Davari, and G. Shahidi, "A 7.9/5.5 psec room/low temperature SOI CMOS," in *Proc. Int. Electron Devices Meeting,* 1997, p. 415.
[93] C. Wann, F. Assaderaghi, L. Shi, K. Chan, S. Cohen, H. Hovel,

TSMC Exhibit 1029

K. Jenkins, Y. Lee, D. Sadana, R. Viswanathan, S. Wind, and Y. Taur, "High-performance 0.07-μm CMOS with 9.5-ps gate delay and 150 GHz f_t," *IEEE Electron Device Lett.,* vol. 18, p. 625, Dec. 1997.

[94] L. T. Su, H. Hu, J. B. Jacobs, M. Sherony, A. Wei, and D. A. Antoniadis, "Tradeoffs of current drive versus short-channel effect in deep-submicrometer bulk and SOI MOSFET's," in *Proc. Int. Electron Devices Meeting,* 1994, p. 649.

[95] S. Biesemans, S. Kubicek, and K. D. Meyer, "Analytical calculations of a figure of merit for novel MOSFET architecture's for the sub 0.25 μm range," *NUPAD,* p. 11, 1994.

[96] C. Fiegna, H. Iwai, T. Wada, M. Saito, E. Sangiorgi, and B. Riccó, "Scaling the MOS transistor below 0.1 μm: Methodology, device structures, and technology requirements," *IEEE Trans. Electron Devices,* vol. 41, p. 941, 1994.

[97] C. Wann, F. Assaderaghi, R. Dennard, C. Hu, G. Shahidi, and Y. Taur, "Channel profile optimization and device design for low-power high-performance dynamic-threshold MOSFET," in *Proc. Int. Electron Devices Meeting,* 1996, p. 113.

[98] Y. Kado, H. Inokawa, Y. Okazaki, T. Tsuchiya, Y. Kawai, M. Sato, Y. Sakakibara, S. Nakayama, H. Yamada, M. Kitamura, S. Nakashima, K. Nishimura, S. Date, M. Ino, K. Takeya, and T. Sakai, "Substantial advantages of fully-depleted CMOS/SIMOX devices as low-power high-performance VLSI components compared with its bulk-CMOS counterpart," in *Proc. Int. Electron Devices Meeting,* 1995, p. 635.

[99] L. Su, M. J. Shernoy, H. Hu, J. E. Chung, and D. A. Antoniadis, "Optimization of series resistance in sub-0.2 μm SOI MOSFET's," *IEEE Electron Device Letters,* vol. 14, p. 363, 1994.

[100] M. Cao, T. Kamins, P. V. Voorde, C. Diaz, and W. Greene, "0.18 μm fully-depleted silicon-on-insulator MOSFET's," *IEEE Electron Device Lett.,* vol. 18, p. 251, 1997.

[101] T. Ando, A. Fowler, and F. Stern, "Electronic properties of two-dimensional systems," *Rev. Modern Phys.,* vol. 54, p. 437, 1982.

[102] S.-I. Takagi, J. Koga, and A. Toriumi, "Subband structure engineering for performance enhancement of Si MOSFET's," in *Proc. Int. Electron Devices Meeting,* 1997, p. 219.

[103] J.-H. Choi, Y.-J. Park, and H.-S. Min, "Electron mobility behavior in extremely thin SOI MOSFET's," *IEEE Electron Device Lett.,* vol. 16, p. 527, 1995.

[104] I. Yang, C. Vieri, A. Chandrakasan, and D. Antoniadis, "Back-gated CMOS on SOIAS for dynamic threshold voltage control," *IEEE Trans. Electron Devices,* vol. 44, p. 822, 1997.

[105] J. Watt and J. Plummer, "Universal mobility-field curves for electrons and holes in MOS inversion layers," in *Proc. Symp. VLSI Technology,* 1987, p. 81.

[106] S. Takagi, I. Iwase, and A. Toriumi, "On the universality of inversion-layer mobility in n- and p-channel MOSFET's," in *Proc. Int. Electron Devices Meeting,* 1988, p. 398.

[107] C.-L. Huang, H. Soleimani, G. Grula, N. Arora, and D. Antoniadis, "Isolation process dependence of channel mobility in thin-film SOI devices," *IEEE Electron Device Lett.,* vol. 17, p. 291, 1996.

[108] S. Tiwari, M. Fischetti, P. Mooney, and J. Welser, "Hole mobility improvement in silicon-on-insulator and bulk silicon transistors using local strain," in *Proc. Int. Electron Devices Meeting,* 1997, p. 939.

[109] M. Horiuchi, T. Teshima, K. Tokumasu, and K. Yamaguchi, "High-current, small parasitic capacitance MOSFET on a poly-Si interlayered (PSI) SOI wafer," in *Proc. Symp. VLSI Technology,* 1995, p. 33.

[110] T. Kachi, T. Kaga, S. Wakahara, and D. Hisamoto, "Variable threshold-voltage SOI CMOSFET's with implanted back-gate electrodes for power-managed low-power and high-speed sub-1-V ulsis," in *Proc. Symp. VLSI Technology,* 1996, p. 124.

[111] C. Wann, K. Noda, T. Tanaka, M. Yoshida, and C. Hu, , *IEEE Trans. Electron Devices,* vol. 43, p. 1742, Oct. 1996.

[112] E. Nowak, J. Johnson, D. Hoyniak, and J. Thygesen, "Fundamental MOSFET short-channel-V_t/saturation current/body effect trade-off," in *Proc. Int. Electron Devices Meeting,* 1994.

[113] H.-S. Wong, K. Chan, and Y. Taur, "Self-aligned (top and bottom) double-gate MOSFET with a 25 nm thick silicon channel," in *Proc. Int. Electron Devices Meeting,* 1997, p. 427.

[114] S. Nakajima, K. Miura, T. Somatani, and E. Arai, "A trench MOSFET with surface source/drain contacts," in *Proc. Int. Electron Devices Meeting,* 1985, p. 200.

[115] D. Hisamoto, T. Kaga, Y. Kamamoto, and E. Takeda, "A fully depleted lean-channel transistor (DELTA)—A novel vertical ultrathin SOI MOSFET," in *Proc. Int. Electron Devices Meeting,* 1989, p. 833.

[116] A.-S. Chu, S. H. Zaidi, and S. Brueck, "Fabrication and raman scattering studies of one-dimensional nanometer structures in (110) silicon," *Appl. Phys. Lett.,* p. 905, 1993.

[117] H.-S. Wong, D. Frank, Y. Taur, and J. Stork, "Design and performance considerations for sub-0.1 μm double-gate SOI MOSFET's," in *Proc. Int. Electron Devices Meeting,* 1994, p. 747.

[118] N. Kistler and J. Woo, "Symmetric CMOS in fully depleted silicon-on-insulator using p^+-polycrystalline Si–Ge gate electrodes," in *Proc. Int. Electron Devices Meeting,* 1993, p. 727.

[119] T.-J. King, J. R. Pfiester, and K. Saraswat, "A variable-work-function polycrystalline-$Si_{l-x}Ge_x$ gate material for submicrometer CMOS technologies," *IEEE Electron Device Lett.,* vol. 12, p. 533, 1993.

[120] C. Kittel, *Introduction to Solid State Physics*. New York: Wiley, 1956, ch. 11, p. 283.

[121] Y. Omura, S. Horiguchi, M. Tabe, and K. Kishi, "Quantum-mechanical effects on the threshold voltage of ultrathin-SOI NMOSFET's," *IEEE Electron Device Lett.,* vol. 14, p. 569, 1997.

[122] Y. Omura, "Features of ultimately miniaturized MOSFET's/SOI: A new stage in device physics and design concepts," *IEICE Trans. Electron.,* vol. E80-C, p. 394, 1997.

[123] H.-S. Wong, D. Frank, and P. Solomon, "Device design considerations for double-gate, ground-plane, and single-gated ultra-thin SOI MOSFET's at the 25 nm channel length generation," in *Proc. Int. Electron Devices Meeting,* 1998, p. 407.

[124] T. Tanaka, H. Horie, S. Ando, and S. Hijiya, "Analysis of p^+ poly Si double-gate thin-film SOI MOSFET's," in *Proc. Int. Electron Devices Meeting,* 1991, p. 683.

[125] J. Denton and G. Neudeck, "Fully depleted dual-gated thin-film SOI p-MOSFET's fabricated in SOI islands with an isolated polysilicon backgate," *IEEE Electron Device Lett.,* vol. 17, p. 509, 1996.

[126] T. Tanaka, K. Suzuki, H. Horie, and T. Sugii, "Ultrafast low-power operation of p^+-n^+ double-gate SOI MOSFET's," in *Proc. Symp. VLSI Technology,* 1994, p. 11.

[127] T. Sugii, T. Tanaka, H. Horie, and K. Suzuki, "15 ps cryogenic operation of 0.19 μm-l_g n^+-p^+ SOI CMOS," on *Proc. SPIE,* vol. 2636, 1995, pp. 74–82.

[128] C. Auth, "Scaling theory for cylindrical fully-depleted, surrounding-gate MOSFET's," *IEEE Electron Device Lett.,* vol. 18, p. 74, 1997.

[129] J. Colinge, M. Gao, A. Romano-Rodriguez, H. Maes, and C. Claeys, "Silicon-on-insulator 'gate-all-around device'," in *Proc. Int. Electron Devices Meeting,* 1990, p. 595.

[130] J.-P. Colinge, "Recent advances in SOI technology," in *Proc. Int. Electron Devices Meeting,* 1994, p. 817.

[131] E. Leobandung, J. Gu, L. Guo, and S. Chou, "Wire-channel and wrap-around-gate metal-oxide-semconductor field-effect transistors with significant reduction in short-channel effects," *J. Vaccuum Sci. Technol.,* vol. B-15, p. 2791, 1997.

[132] P. Francis, A. Terao, D. Flandre, and F. V. de Wiele, "Characteristics of nMOS/GAA (gate-all-around) transistors near threshold," *Microelectron. Eng.,* vol. 19, p. 815, 1992.

[133] F. Assaderaghi, S. Park, D. Sinitsky, J. Bokor, P.-K. Ko, and C. Hu, "A dynamic threshold voltage MOSFET (DTMOS) for low voltage operation," *IEEE Electron Device Lett.,* vol. 15, p. 510, 1994.

[134] I.-Y. Chung, Y.-J. Park, and H.-S. Min, "A new SOI inverter using dynamic threshold for low-power applications," *IEEE Electron Device Lett.,* vol. 18, p. 248, 1997.

[135] R. Zingg and B. Höfflinger, "Stacked CMOS inverter with symmetric device performance," in *Proc. Int. Electron Devices Meeting,* 1989, p. 909.

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

TSMC Exhibit 1029

[137] K. Jenkins and J. Sun, "Measurement of I–V curves of silicon on insulator (SOI) MOSFET's without self heating," *IEEE Electron Device Lett.,* vol. 16, p. 145, 1995.
[138] L. Su, J. Chung, A. Antoniadis, K. Goodson, and M. Flik, "Measurement and modeling of self heating in SOI nMOSFET's," *IEEE Trans. Electron Devices,* vol. 41, p. 69, 1994.
[139] L. Guo, E. Leobandung, and S. Chou, "A room-temperature silicon single-electron metal-oxide-semiconductor memory with nanoscale floating-gate and ultranarrow channel," *Appl. Phys. Lett.,* vol. 70, no. 7, p. 850, 1997.
[140] A. Nakajima, T. Futatsugi, K. Kosemura, T. Pukano, and N. Yokoyama, "Room temperature operation of Si single-electron memory with self-aligned floating dot gate," *Appl. Phys. Lett.,* vol. 70, no. 13, p. 1742, 1997.
[141] J. Welser, S. Tiwari, S. Rishton, K. Lee, and Y. Lee, "Room temperature operation of a quantum-dot flash memory," *IEEE Electron Device Lett.,* vol. 18, no. 6, p. 278, 1997.
[142] S. Tiwari, F. Rana, K. Chan, H. Hanafi, W. Chen, and D. Buchanan, "Volatile and nonvolatile memories in silicon with nano-crystal storage," in *Proc. Int. Electron Devices Meeting,* 1995, p. 521.
[143] K. Yano, T. Ishii, T. Hashimoto, T. Kobayashi, F. Murai, and K. Seki, "Room-temperature single-electron memory," *IEEE Trans. Electron Devices,* vol. 41, p. 1628, Sept. 1994.
[144] K. Yano, T. Ishii, T. Mine, F. Murai, and K. Seki, "Single-electron-memory integrated circuit for giga-to-tera bit storage," in *Proc. Int. Solid State Circuits Conf.,* 1996, p. 266.
[145] K. Yano, T. Ishii, T. Sano, T. Mine, F. Murai, T. Kure, and K. Seki, "A 128 Mb early prototype for gigascale single-electron memories," in *Proc. Int. Solid State Circuits Conf.,* 1998, p. 344.
[146] T. Ishii, K. Yano, T. Sano, T. Mine, F. Murai, and K. Seki, "Verify: Key to the stable single-electron-memory operation," in *Proc. Int. Electron Devices Meeting,* 1997, p. 171.
[147] S. Tiwari, J. Welser, and F. Rana, "Technology and power-speed trade-offs in quantum-dot and nano-crystal memory devices," in *Proc. Symp. VLSI Technology,* 1997, p. 133.
[148] S. Tiwari, F. Rana, H. Hanafi, E. Crabbé, and K. Chan, "A silicon nanocrystals based memory," *Appl. Phys. Lett.,* vol. 68, no. 10, p. 1377, 1996.
[149] C. H.-J. Wann and C. Hu, "High endurance ultrathin tunnel oxide for dynamic memory applications," in *Proc. Int. Electron Devices Meeting,* 1995, p. 867.
[150] K. Yano, T. Ishii, T. Sano, T. Mine, F. Murai, and K. Seki, "Impact of Coulomb blockade on low-charge limit of memory device," in *Proc. Int. Electron Devices Meeting,* 1995, p. 525.
[151] T. Ishii, K. Yano, T. Sano, T. Mine, F. Murai, T. Kure, and K. Seki, "A 3-D single-electron-memory cell structure with 2 F^2 per bit," in *Proc. Int. Electron Devices Meeting,* 1997, p. 924.
[152] B. Furht, "A survey of multimedia compression techniques and standards—Part II: Video compression," *Real-Time Imaging,* vol. 1, pp. 319–337, 1995.
[153] B. Ackland, "Video compression and VLSI," in *Proc. Custom Integrated Circuits Conf.,* 1993, pp. 11.1.1–6.
[154] W. C. Jakes, *Microwave Mobile Communications.* Piscataway, NJ: IEEE Press, 1994.
[155] A. K. Lenstra and H. W. L., Jr., "The development of the number field sieve," in *Lecture Notes in Math*, vol. 1554. Berlin, Germany: Springer-Verlag, 1993.
[156] T. L. Sterling, *Enabling Technologies for Petaflops Computing.* Cambridge, MA: MIT Press, 1995.
[157] C. Wann, L. Su, K. Kenkins, R. Chang, D. Frank, and Y. Taur, "RF perspectives of sub-tenth-micron CMOS," in *Proc. Int. Solid State Circuits Conf.,* 1998, p. 254.
[158] T. Seshita *et al.,* "A 20 GHz 8 bit multiplexer IC implemented with 0.5 μm wnx/w-gate GAAS MESFET," *IEEE J. Solid State Circuits,* vol. 29, pp. 1583–1588, Dec. 1994.
[159] R. Yan *et al., IEEE Electron Device Lett.,* 1992.
[160] E. Crabbe *et al., IEEE Electron Device Letters,* 1992.
[161] E. R. Fossum, "Active pixel sensors: Are CCD dinosaurs?," *Proc. SPIE,* vol. 1900, p. 2, 1993.
[162] ——, "CMOS image sensors: Electronic camera on a chip," in *Proc. Int. Electron Devices Meeting,* 1995, p. 17.
[163] B. Ackland and A. Dickinson, "Camera on a chip," in *Proc. Int. Solid State Circuits Conf.,* 1996, p. 22.
[164] H.-S. Wong, "Technology and device scaling considerations for CMOS imagers," *IEEE Trans. Electron Devices,* vol. 17, pp. 2131–2142, Dec. 1996.
[165] H.-S. P. Wong, "CMOS image sensors—Recent advances and device scaling considerations," in *Proc. Int. Electron Devices Meeting,* 1997, p. 201.
[166] J. Bosiers, Y. Boersma, A. Kleinmann, D. Verbugt, H. Peek, and A. van dr Sijde, "A 1/3 inch progressive scan 1280(H) × 960 (V) FT-CCD for digital still camera applications," in *Proc. Int. Electron Devices Meeting,* 1997, p. 185.
[167] H. Ihara, H. Yamashita, I. Inoue, T. Yamaguchi, N. Nakamura, and H. Nozaki, "A 3.7 μm × 3.7 μm, square pixel CMOS image sensor for digital still camera application," in *Proc. Int. Solid State Circuits Conf.,* 1998, p. 182.
[168] B. Fowler, A. E. Gamal, and D. X. D. Yang, "A CMOS area image sensor with pixel-level A/D conversion," in *Proc. Int. Solid State Circuits Conf.,* 1994, p. 226.
[169] S. Smith, J. Hurwitz, M. Torrie, D. Baxter, A. Holmes, M. Panaghiston, R. Henderson, A. Murray, S. Anderson, and P. Denyer, "A single-chip 306 × 244 pixel CMOS NTSC video camera," in *Proc. Int. Solid State Circuits Conf.,* 1998, p. 170.
[170] M. Loinaz, K. Singh, A. Blanksby, D. Inglis, K. Azadet, and B. Ackland, "A 200-mW 3.3 V CMOS color camera IC producing 352 × 288 24-b video at 30 frames/s," in *Proc. Int. Solid State Circuits Conf.,* 1998, p. 168.
[171] J. Hertz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation* (Lecture Notes Volume I: Santa Fe Institute Studies in the Sciences of Complexity). Reading, MA: Addison-Wesley, 1991.
[172] C. Sah, "Evolution of the MOS transistor—From conception to VLSI," *Proc. IEEE,* pp. 1280–1326, 1988.
[173] R. B. Fair, "History of some early developments in ion-implantation technology leading to silicon transistor manufacturing," *Proc. IEEE,* vol. 86, p. 111, Jan. 1998.
[174] P. Bondy, "Moore's law governs the silicon revolution," *Proc. IEEE,* vol. 86, p. 78, Jan. 1998.
[175] B. Crowder and S. Zirinsky, "Metal silicide interconnection technology—A future perspective," *IEEE Trans. Electron Devices,* vol. 26, p. 369, 1979.
[176] C. Lau, Y. See, D. Scott, J. Bridges, S. Perna, and R. Davies, "Titanium disilicide self-aligned source/drain + gate technology," in *Proc. Int. Electron Devices Meeting,* 1982, p. 714.
[177] R. Rung, H. Momose, and Y. Nagakubo, "Deep trench isolated CMOS devices," in *Proc. Int. Electron Devices Meeting,* 1982, p. 237.
[178] S. Wong, C. Sodini, T. Eckstedt, H. Grinolds, K. Jackson, and S. Kwan, "Low pressure nitrided oxide as a thin gate dielectric for MOSFET's," *J. Electrochem. Soc.,* vol. 130, p. 1139, 1983.
[179] C. Codella and S. Ogura, "Halo doping effects in submicron di-ldd device design," in *Proc. Int. Electron Devices Meeting,* 1985, p. 230.
[180] J. Y.-C. Sun, Y. Taur, R. Dennard, S. Klepner, and L. Wang, "0.5 μm-channel CMOS technology optimized for liquid-nitrogen-temperature operation," in *Proc. Int. Electron Devices Meeting,* 1986, p. 236.
[181] S. Hillenius, R. Liu, G. Georgiou, D. W. R. L. Field, A. Kornblit, D. Boulin, R. Johnston, and W. Lynch, "A symmetric submicron CMOS technology," in *Proc. Int. Electron Devices Meeting,* 1986, p. 252.
[182] B. Davari, C. Koburger, R. Schulz, M. J. J. D. Warnock, Y. Taur, W. Schwittek, M. K. J. K. DeBrosse, and J. Mauer, "A new planarization technique, using a combination of RIE and chemical mechanical polish (CMP)," in *Proc. Int. Electron Devices Meeting,* 1989, p. 61.
[183] F. White, W. Hill, S. Eslinger, E. Payne, W. Cote, B. Chen, and K. Johnson, "Damascene stud local interconnect in CMOS technology," in *Proc. Int. Electron Devices Meeting,* 1992, p. 301.
[184] J. Paraszczak, D. Edelstein, S. Cohen, E. Babich, and J. Hummel, "High performance dielectrics and processes for ULSI interconnection technologies," in *Proc. Int. Electron Devices Meeting,* 1993, p. 261.

TSMC Exhibit 1029

Hon-Sum Philip Wong (Senior Member, IEEE) received the B.Sc. (Hon.) degree from University of Hong Kong, Hong Kong, in 1982 and the Ph.D. degree in electrical engineering from Lehigh University, Bethlehem, PA, in 1988.

He joined IBM T. J. Watson Research Center, Yorktown Heights, NY, in 1988 as a Research Staff Member. From 1988 to 1992, he worked on the design, fabrication, and characterization of a high-resolution, high color-fidelity CCD image scanner for art work archiving. Since 1993, he has been working on analysis, fabrication, and applications of nanoscale CMOS devices. His recent work included simulations of discrete random dopant fluctuation effects in small MOSFET's, the physics and technology of double-gate and back-gate MOSFET's, CMOS projection displays, and CMOS image sensors.

David J. Frank (Member, IEEE) received the B.S. degree from the California Institute of Technology, Pasadena, in 1977 and the Ph.D. degree in physics from Harvard University, Cambridge, MA, in 1983.

Since 1983, he has been employed at the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he is a Research Staff Member. His studies have included nonequilibrium superconductivity, modeling and measuring III-V devices, and exploring the limits of scaling of silicon technology. His recent work includes the modeling of innovative silicon devices, analysis of CMOS scaling issues, investigating the usefulness of energy-recovering CMOS logic and reversible computing concepts, and low-power circuit design. His interests include superconductor and semiconductor device physics, modeling and measurement, circuit design, and percolation in two-dimensional systems.

Paul M. Solomon (Fellow, IEEE) was born Cape Town, South Africa. He received the B.Sc. degree in electrical engineering from the University of Cape Town, South Africa, in 1968 and the Ph.D. degree from the Technion, Haifa, Israel, in 1974 for work on the breakdown properties of silicon dioxide.

Since 1975, he has been a Research Staff Member at the IBM T. J. Watson Research Center, Yorktown Heights, NY. At IBM, his interests have been in the field of high-speed semiconductor devices. He has contributed to the theory of scaling bipolar transistors to very small dimensions and has developed methodologies to compare the performance of high-speed semiconductor devices. The design of high-speed semiconductor logic devices has been a continuing topic, ranging from self-aligned bipolar transistors through novel heterostructure field effect transistors and more recently to novel CMOS device concepts. He has contributed to the physics of transport in semiconductors and has taught the physics of high-speed devices at Stanford University.

Dr. Solomon is a member of APS.

Clement H. J. Wann received the B.S. degree from National Taiwan University in 1988 and the M.S. and Ph.D. degrees from University of California, Berkeley, in 1992 and 1996, respectively, all in electrical engineering.

He joined IBM T. J. Watson Research Center, Yorktown Heights, NY, as a Research Staff Member in 1996. He is currently with IBM Semiconductor Research and Development Center, East Fishkill, NY.

Jeffrey J. Welser (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1988, 1989, and 1994, respectively.

He has held short-term positions at Sumitomo Electric in Japan (1988) and at IBM T. J. Watson Research Center (1989–1990), Yorktown Heights, NY, working on GaAs devices, and a postdoctoral position at Stanford University (1995), where he continued his thesis research on SiGe materials and their applications to MOSFET devices. Since 1995, he has been a Research Staff Member at IBM T. J. Watson Research Center, and his current research activities focus on novel silicon devices, including vertical transistors and nanostructures, for a variety of memory applications.

TSMC Exhibit 1029