SOI CMOS as a Mainstream Low-Power Technology: A Critical Assessment

Dimitri A. Antoniadis

Microsystems Technology Laboratories
Massachusets Institute of Technology
Cambridge, MA 02139, USA

Abstract

This paper provides an overview of SOI MOSFET theory and practice with emphasis on circuit applications issues. Fully and partially depleted channel devices are considered and particular attention is given to describing the so-called floating-body effects that are unique to SOI. Two advanced SOI MOSFET configurations, dual-gate SOI and active-body SOI, specifically developed for low voltage circuit applications are also discussed.

Introduction

Silicon on insulator, SOI, is considered advantageous with respect to bulk silicon as a substrate for CMOS technology, particularly for low power applications. Figure 1 shows an idealized cross-section through the two transistors of an SOI CMOS inverter. As can be seen the transistors are entirely isolated from each other with their active areas completely surrounded by silicon dioxide insulator. Being dielectrically isolated from each other CMOS devices are not subject to CMOS latch-up. In addition, their channel area, so-called *body*, is normally floating electrically unless it is specifically contacted as a forth terminal, or is shorted to the source. Floating-body implementation gives the smallest device area and is therefore prefered, therefore it is the key distiguishing feature between bulk and SOI MOSFET operation.

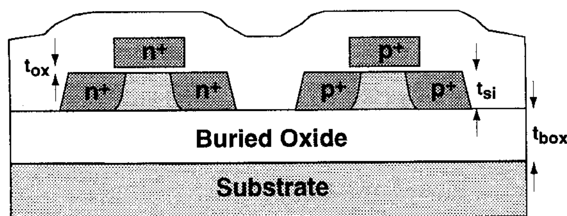


Figure 1. Idealized cross section through two SOI CMOS transistors. Three important film thicknesses are defined: gate oxide thickness, t_{ox}, silicon film thickness, t_{si}, and buried oxide thickness, t_{box}.

SOI MOSFETs are broadly classified, depending on the depletion condition of their channel body during operation, into fully depleted (FD) and partially depleted (PD) MOSFETs. While all other SOI features, i.e. source/drain capacitance and layout density, are the same FD and PD devices can exhibit significant operational differences. Also, they have different manufacturability issues. These are discussed in the following sections.

There are three advantages of SOI MOSFETs relative to bulk MOSFETs for low power circuit applications: 1) They exhibit significantly reduced source- and drain-to-substrate capacitance because of the elimination of the standard junction of bulk MOSFETs. Depending on the reference bulk technology as much as tenfold capacitance reduction is achieved. 2) Because of their dielectric isolation they do not exhibit the conventional *body-effect* that decreases current drive in stacked devices (e.g. NMOS in a NAND gate), particularly with scaled- down power supply voltage, V_{dd}. 3) When their body floats it couples to the gate potential resulting in higher ON/OFF current ratio than their bulk counterparts, and therefore they are more suitable for reduced V_{dd} operation at given performance. However, as we will see, this floating-body feature makes SOI CMOS device and circuit engineering more complex than for bulk.

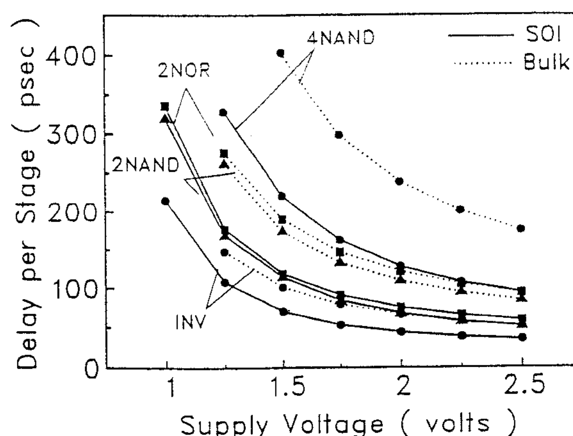


Figure 2. Delay per stage vs. power supply voltage for inverter, 2-input NOR, 2-input NAND, and 4-input NAND stages, implemented in bulk and SOI CMOS with approximately equal channel lengths and threshold voltages. [After Shahidi in "A Tutorial on SOI Materials Devices and Technologies", 1995 Int. SOI Conf. Short Course.]

Figure 2 illustrates these advantages by comparing delay per stage between SOI and bulk CMOS with aproximately the same channel length and threshold voltage. Simple inverter, 2-input NOR, 2-input NAND, and 4-input NAND stages are compared. At V_{dd}=1.5 V the relative delay reduction in SOI is 35%, 59%, 63%, and 85%,

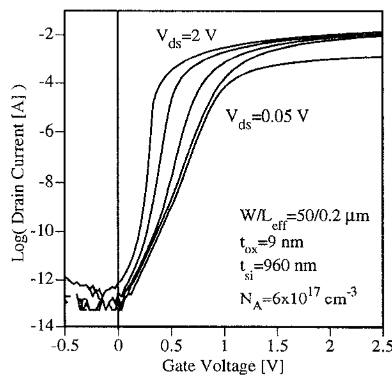
Permission to make digital/hard copy of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of ACM, Inc. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
©1997 ACM 0-89791-903-3/97/08..$3.50

respectively. The delay reduction in the inverter case partitions approximately, 25% due to capacitance reduction and 10% due to gate-body coupling. The larger delay reduction in the other stages is due to the body-effect improvement of stacked devices. Since power dissipation is approximately CV_{dd}^2 the reduced nodal capacitance, C, leads to 25% power reduction *at constant* V_{dd}. However, if delay decrease were to be traded-off for power reduction, e.g. the bulk would have to run at V_{dd}=2.15 V to achieve same delay as SOI at 1.5 V, then the relative power in SOI is $0.75x(1.5/2.15)^2$=0.36, i.e. power is reduced by a factor of 2.7 at *constant delay.*

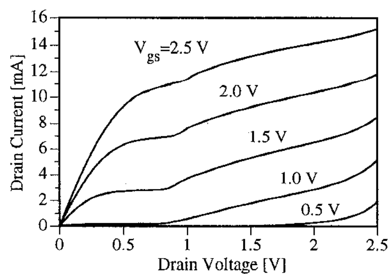
In addition to its application in straightforward CMOS, SOI also lends itself to the realization of novel device structures that are not possible in bulk silicon wafers. Such device structures of interest to low power applications are the dual gate SOI MOSFET which incorporates a gate under the silicon film, and the active-body MOSFET in which the channel body voltage is externally controlled. As will be seen later both of these device structures allow dynamic control of the MOSFET threshold voltage and hence even more aggressive power supply voltage scaling than simple SOI with concomitant power savings.

Partially Depleted SOI MOSFETs

CMOS in PD-SOI is easier to manufacture than in FD-SOI. On the other hand, unless the channel body is



(a)



(b)

Figure 3. (a) Static transfer characteristics, $\log I_d$ vs. V_{gs}, of PD-SOI nMOSFET. (b) Static output characteristics, I_d vs. V_{ds}.

intentionally tied electrically to the source [e.g. 2], they exhibit marked *floating-body effects.* Figure 3 shows typical transfer and output *static I-V* characteristics of a PD nMOSFET. Figure 4 depicts a cross section of the device and illustrates the processes taking place at some bias condition $V_{gs}>V_{th}$ and $V_{ds}>0$. The reason for the two unusual I-V features, the increasing "shoulder" in the transfer characteristics, and the "kink" in the output characteristics is the varying body voltage, V_{bs}, which affects the device V_{th}. V_{bs}, under static conditions is established by the balance between impact ionization current and diode leakage to the source. However, it has been shown that these *static floating-body bipolar effects* (*bipolar* because carriers of both polarities are involved) are absent under normal, rapid switching operation because the impact ionization current is too weak to rapidly change the body charge, and hence affect V_{bs} [3]. Therefore, the apparent super-steep subthreshold slope, e.g. in Fig. 3, cannot be taken advantage of for low voltage operation.

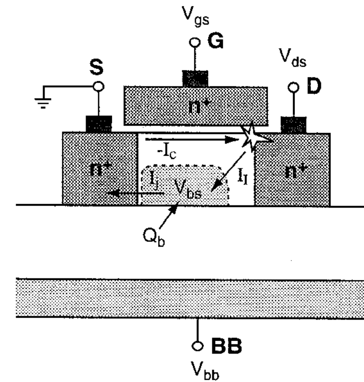


Figure 4. Illustration of typical current flows in a PD-SOI nMOSFET at $V_{gs}>V_{th}$, $V_{ds}>0$. Channel current, I_c (electrons), impact ionization current, I_I (holes), and body-source diode current, I_j (electron-hole recombination current

Of more significance for circuit operation are the *dynamic floating-body effects* which can be distinguished in two categories, *capacitive,* and *bipolar turn-on.* These effects affect the instantaneous value of V_{bs} and hence of the drain current. *Capacitive* effects arise from the capacitive coupling of the floating-body to the gate, source and drain. Contrary to bulk where body charge, Q_b, is free to move in the form of majority carrier current to or from the *fixed-voltage* body in response to gate, source, and drain voltages, in PD-SOI MOSFETs this charge is near-constant in the time-scale of typical logic clock frequencies and therefore V_{bs} has to *vary* in that time-scale [4,5].

While the slow change of Q_b causes a slow change in the *average value* of V_{bs}, direct capacitive coupling to the gate and drain voltages during transients causes rapid V_{bs} fluctuations. The slow Q_b change can be a concern when a circuit starts switching from an idling state. Effectively, V_{th} would vary in the course of many clock cycles rising or

decreasing toward its *switching-steady-state* value, *depending on* the device initial state [6,7].

On the other hand, the rapid V_{bs} fluctuation will lead to a more or less repetitive modulation of the device current relative to fixed V_{bs}. The gate voltage ramp couples to V_{bs} producing a positive spike that increases the dynamic subthreshold slope (not shown), a beneficial effect. Thus, the net effect of *capacitive dynamic floating-body effects* is to improve the rate of device turn on, and introduce a slowly hysteretic V_{th}. Proper device engineering can minimize the latter, while maintaining the first.

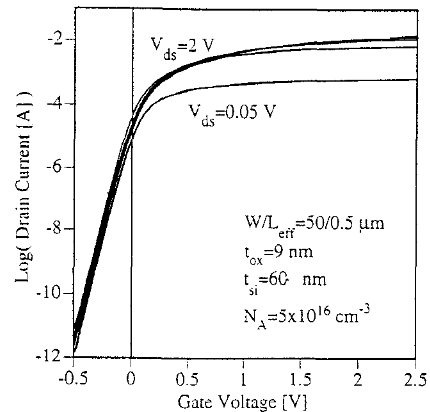
Dynamic bipolar turn-on floating-body effects can arise under conditions that result in V_{bs} sufficiently high to turn-on the source diode. Then the parasitic npn bipolar transistor formed by the source-body-drain turns on and bipolar current flows from drain to source [8,9]. This brief bipolar turn-on and current flow can occur under some pass-transistor conditions. Fortunately, the total injected charge is quite small, and the set-up time required for this effect is long enough so that the effect is easily manageable [8].

Summarizing, PD-SOI CMOS can be built on a variety of different silicon film thickness. The key disadvantage for this device type comes from the floating-body effects which need to be carefully managed. Once understood, devices can be optimized to reduce the severity of these effects, with what remains of them manageable at the circuit design level. Also, floating-body effects can be eliminated by tying the quasi-neutral body of the PD-SOI MOSFET to the source. However, this results in increased area per transistor, as well as electrical asymmetry with respect to source and drain. Therefore this solution is reserved for special circumstances in circuit design.

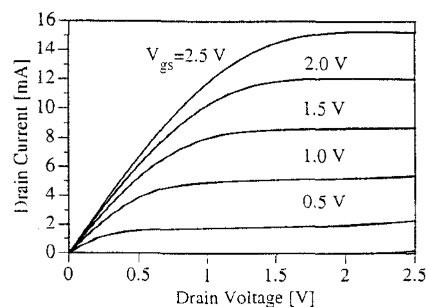
Fully Depleted SOI MOSFETs

When the SOI silicon film thickness, t_{si}, is less than the depletion depth, t_d, under the inversion channel, SOI MOSFETs are said to be fully depleted. In a FD-SOI MOSFET the body charge cannot be modulated by the terminal voltages. Also, because the whole body is depleted the diode potential barrier is decreased such that the required forward bias for significant conduction can be much smaller than the usual 0.7 V. Thus, any injected impact ionization current from the drain is shunted to the source, and thus V_b remains near 0 V. Therefore PD-SOI MOSFETs should exhibit much reduced floating-body effects. Figure 5 shows an example of FD-SOI nMOSFET transfer and output static I-V characteristics. Note the absence of kinks and shoulders, in contrast to Fig. 3.

From the low power application standpoint another advantage of FD SOI is that the subthreshold slope, 2.3nkT/q volts per decade of current, can approach ideal where n=1. This leads to maximal on/off current ratio



(a)



(b)

Figure 5. (a) Static transfer characteristics, $\log I_d$ vs. V_{gs}, of FD-SOI nMOSFET. (b) Static output characteristics, I_d vs. V_{ds}.

which allows minimization of V_{dd}, and hence power, for given I_{off} and performance.

Since the body is fully depleted it contributes a fixed charge which determines the device V_{th}. Therefore, depending on the channel doping configuration, the body charge, and hence V_{th}, can depend on the silicon thickness which together with its required thinness places an increased manufacturability burden on FD vs. PD CMOS. While, a doping scheme has been demonstrated that reduces the FD-CMOS V_{th} sensitivity to t_{si} variation [11], the required silicon thickness for scaled FD devices which is less than 50 nm, poses a significant challenge to the formation a low resistance contacts to source and drain [12].

As one can imagine, the transition from PD ($t_{si}>t_d$) to FD ($t_{si}<t_d$) behavior is not abrupt. First, it should be noted that since the depletion depth normally increases with distance from the source to the drain, the relevant depletion depth here is the one near the source, i.e. the minimum depletion depth in the channel. Second, for given doping and t_{si} that produce a marginally PD *long channel* MOSFET, it is quite possible that *short channel* devices are FD because of

increased body depletion provided by the source and drain due to the well-known *charge sharing effect*. Finally, the diode barrier can be expected to decrease slowly from its normal value (V_{diode} ~0.7 V) as t_{si} decreases below t_d. It is clear then that the degree of floating-body effect amelioration would increase as t_{si} decreases below t_d. This effect is illustrated in Fig. 6 which plots three transient values of V_{th}, vs. t_{si} required to achieve nMOSFETs with constant I_{off} (1 nA/μm) [13]. These values of V_{th} depend on the previous state of the transistor; switching after long idle with drain at V_{dd}, or after long idle with drain at 0 V, or constantly switching. As can be seen a film thickness of < 25 nm is required for all these V_{th}'s to merge together, indicating complete absence of floating-body effects. This thickness is less than that required for full depletion and is channel length specific.

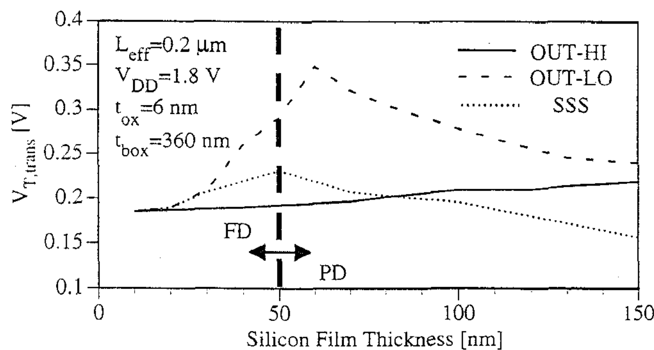


Figure 6. Transient threshold voltage, $V_{T,trans}$, for SOI nMOSFETs with characteristic values shown in the inset, vs. silicon film thickness, t_{si}. Three different values of $V_{T,trans}$ are shown. OUT-HI and OUT-LO are the effective threshold voltages *immediately after* idling at drain high and low states, respectively. After switching steady state (SSS) has been reached (e.g. as in Fig. 4), the effective threshold voltage settles to the SSS values.

Summarizing, while FD-SOI MOSFETs are attractive because of decreased floating-body effects, they are difficult to manufacture because they require very thin tsi in order to be truly free of these effects. For this reason a compromise between severity of floating-body effects and manufacturability dictates tsi values thicker than optimum, e.g. order of 50 nm for 0.25 μm CMOS technology.

Self-Heating Effects

Because SOI MOSFETs are separated from the underlying silicon wafer by a thin layer of SiO_2 they are less coupled thermally to the wafer then their bulk MOSFET counterparts. This leads to the well known *self-heating effect* . This effect is evident in the FD and less so in the PD device characteristics (see Fig. 5 at $V_{gs}$of 2 V and 2.5 V). It reduces the drain current, and in the more severe cases of high V_{gs} and V_{ds} gives rise to negative differential output conductance. It turns out that in logic applications this effect is not as bad as it looks in the static characteristics; the typical power dissipated per switching cycle in the most heavily loaded logic device is approximately a factor of 10 less than the nominal device $I_{dmax}V_{dd}$, and therefore the device temperature increase due to self-heating is typically < 8 degrees [14]. On the other hand, the self-heating effect is also dynamic with time constant of order 1 μs, and therefore cannot respond to the typical logic clock rates. For both of these reasons self-heating is of no concern for logic applications. However, it should be considered carefully in analog applications [15].

Active-Body SOI MOSFETs

Active-body SOI MOSFETs are partially-depleted SOI MOSFETs designed such that their channel body can be externally electrically controlled as a fourth terminal [16]. Essentially, the body becomes a controllable "back-gate" in the same sense as in bulk-Si CMOS wells. Hence their V_{th} can be adjusted by controlling V_{bs}, and this can be taken advantage of in low power CMOS circuit optimization [17, 18].

A special case of active-body SOI, where the body is directly tied to the gate, has also been proposed as an advantageous configuration for low-power circuit applications [19]. This so-called dynamic-threshold, DT-CMOS is only applicable for very low V_{dd}, so that the body to source leakage can be negligible. By tying the gate to the body the DT-CMOS device achieves an ideal subthreshold slope with n=1, which is the same as that of an ideal FD-SOI CMOS, but without the manufacturing difficulties of the very thin t_{si}, discussed in the FD-SOI section. The ideal suthreshold slope gives rise to maximization of on/off current ratio which allows minimum power for given performance, as outlined in the FD-SOI section. Because the body in these devices is tied to the gate, and hence is not floating, they should not exhibit any of the floating-body effects expected in PD-SOI.

While attractive in its manufacturing simplicity, the straigth-forward active-body PD-SOI device configuration in which the body is contacted at the widthwise edge of the channel, is limited to relatively small widths by the R-C time constant of the body back-gate. If the same R-C criterion that applies in determining the maximum channel width between normal front-gate contacts is applied to the body then the maximum width would have to decrease by approximately R_f/f_cR_b, where R_f and R_b are the sheet resistances of front-gate and body, respectively, and f_c is the ratio of the front to back capacitances seeing by the two terminals. For technologies in the 250 nm generation the width decrease would be of order 50-100. Further scaling-down aggravates this problem.

Dual-Gate SOI MOSFETs

Dual-gate SOI MOSFETs are fully-depleted SOI MOSFETs with a gate underneath the channel in addition to the normal top gate. This bottom gate can be either electrically tied directly to the top gate [20, 21], or it can be independently controlled [22]. The objective for the tied top to bottom gate concept is to increase the current drive of the MOSFET by providing two parallel inversion channels from source to drain - the top one induced by the top-gate and the bottom one induced by the bottom-gate. For this to work effectively the two oxide thicknesses must be equal ($t_{ox}=t_{box}$). Precise alignment of top to bottom gates is necessary to avoid unwanted parasitic overlap capacitances to source and drain. While very difficult to realize it, this kind of dual-gate MOSFET can be shown to be the ultimate scaleable MOSFET configuration [23]. From the low-power circuit application standpoint this dual-gate MOSFET can be considered as a PD-SOI MOSFET with twice the current drive per unit area, and hence somewhat more than half the source/drain parasitic capacitance per unit current, depending on the bottom-gate to drain source/drain overlap. However, the total gate capacitance per unit current is the same or slightly increased compared to the single-gate PD-SOI MOSFET.

On the other hand, the independently controlled dual-gate MOSFET is specifically designed such that the bottom-gate can control the V_{th} of the top-gated MOSFET, in a so-called SOIAS (Silicon On Insulator on Active Substrate) configuration [22]. This allows the device threshold to be dynamically adjusted during circuit operation. The main low-power application is in logic circuits with variable activity, e.g. event-driven computation, where V_{th} can be adjusted to be low for high performance during computation, and then adjusted to be high for low leakage during idle periods. Fig. 7 shows an example of dual-gate nMOS transfer I-V characteristics at two values of V_{th}.

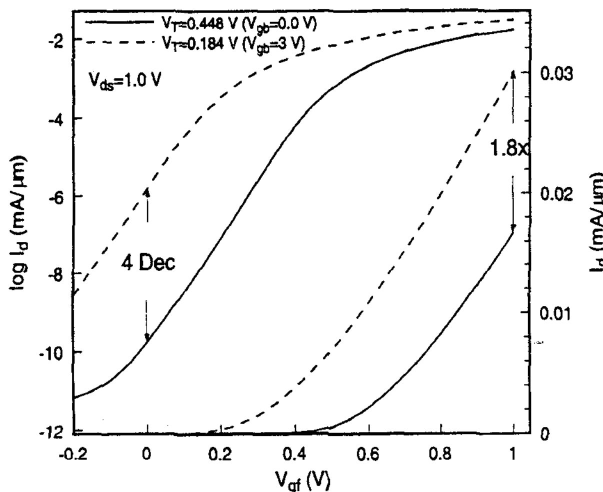


Figure 7. Measured nMOSFET transfer linear and log-linear I-V charcateristics. tuned at different V_{th}. L_{eff}=0.44 microns.

A ratio of $2x10^8$ in active to idle current is achieved at V_{dd}=1 V. Other possible applications of dynamic V_{th}-control include compensation of current-drive-variation due to temperature and manufacturing variations. In all of these kinds of applications groups of devices are likely to have their V_{th} controlled simultaneously, therefore it is very likely that backgate contacts would not be needed for each individual device.

Of course, the variation of V_{th} does cost energy due to the back-gate capacitance and voltage change. Therefore the energy savings from this dynamic V_{th} operation would depend strongly on circuit activity which in turn depends on overall system activity. They would also depend on device design parameters such as t_{box} and parasitic capacitances such as back-gate to source/drain overlap. Fig. 8 shows the ratios of energy expended to perform a given computation in a hypothetical SOIAS to an otherwise identical SOI CMOS technology, for various modules of a hypothetical

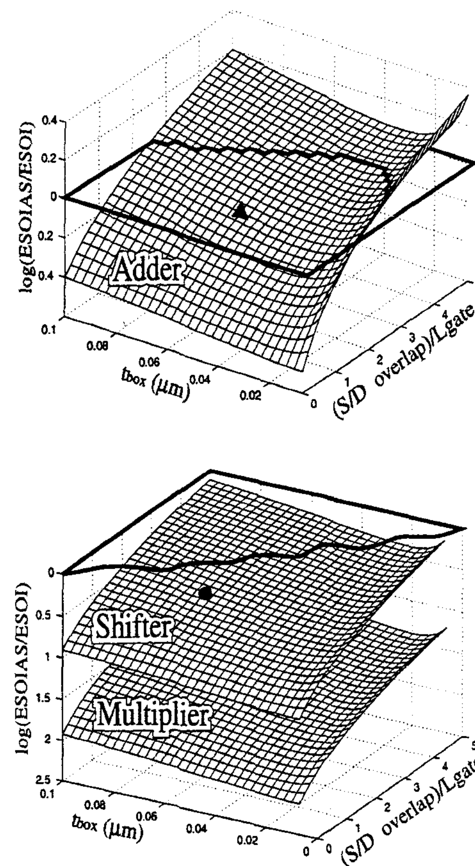


Figure 8. Energy ratio of SOIAS to SOI CMOS in a technology design space for the adder , shifter, and multiplier functional modules operating in 2% duty cycle burst mode. The dark line demarcates the break-even plane. In order to minimize the energy cost of back-gate switching thesource/drain to back-gate overlap must be minimized and the t_{box} must be optimized.

microprocessor operating in 2% burst mode, equivalent to an active x-terminal [24]. The energy savings are largest for the lowest-activity multiplier unit. The dependence on t_{box} is weak because of the tradeoff between back-gate capacitance and required voltage change to effect a given V_{th} change. On the other hand the dependence on back-gate to source and drain overlap is large.

Summary and Conclusions

Standard SOI-CMOS, whether with partially depleted or fully depleted channel body, allows power reduction at given performance relative to bulk CMOS. This power reduction is of order 35%, but the exact amount will depend on the low-power optimization level of the reference bulk technology. PD-SOI is more readily manufacturable than FD-SOI, but it exhibits stronger floating-body effects which must either be eliminated or at least reduced and then taken into consideration in circuit design. The SOI configuration opens up the possibility of novel MOSFET structures, active-body SOI and dual-gate SOI, which can lead to even further reduction of power for given performance by means of new circuit design techniques.

Acknowledgments

The work of my past and current students Lisa Su, Isabel Yang, Melanie Sherony, Andy Wei and Anthony Lochtefeld, as well as continuing support from the SRC and DARPA are gratefully acknowledged.

References

[1] T. Iwamatsu et al., *IEEE Trans. on Elec. Dev.*, vol. 42, no. 11, pp. 1934-1939, 1995.
[2] D. Suh and J. G. Fossum, *IEEE Trans. on Elec. Dev.*, vol. 42, no 4, pp 728-737, 1995.
[3] A. Wei et al., *IEEE Electron Dev. Letters*, vol. 16, no. 11, pp. 494-496, 1995.
[4] A. Wei et al., *IEEE Electron Dev. Letters*, vol. 17 , no. 8, pp. 391-394, 1996 .
[5] J. Gautier et al., *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp.623-626, 1995.
[6] D. Suh and J. G. Fossum, *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 661-664, 1994.
[7] A. Wei et al.,*Proceedings 1996 IEEE International SOI Conference, pp. 74-75, 1996.*
[8] A. Wei et al., *IEEE Electron Dev. Letters*, vol. 17, no. 5, pp. 193-195, 1996.
[9] M. M. Pellela et al., *IEEE Electron Dev. Letters*, vol. 17, no. 5, pp. 196-198, 1996.
[10] H. Lim, and J. G. Fossum, *IEEE Trans. on Elec. Dev.*, vol. 31, pp. 401-408, 1985.
[11] M. J. Sherony et al., *IEEE Electron Dev. Letters*, vol. 16, no. 3, pp. 100-102, 1995.
[12] L. T. Su et al., *IEEE Electron Dev. Letters*, vol. 15, no. 9, pp. 363-365, 1994.
[13] M. J. Sherony et al., *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 125-128, 1996.
[14] L. T. Su et al., *IEEE Trans. on Elec. Dev.*, vol. 41, pp. 69-75, 1994.
[15] B. M. Tenbroek et al., *IEEE Trans. on Elec. Dev.*, vol. 43, pp. 2227-2232, 1985.
[16] J. P. Colinge, Kluwer Academic Publishers, Boston, MA, 1991.
[17] T. Douseki et al., *Int. Solid-State Circuits Conf. Tech. Dig., pp 84-85, 1996.*
[18] T. Fuse et al., *Int. Solid-State Circuits Conf. Tech. Dig., pp 88-89, 1996.*
[19] F. Assaderaghi et al., *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 809-812, 1994.
[20] J. P. Colinge et al., *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 595-598, 1990.
[21] T. Tanaka et al., *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 683-686, 1991.
[22] I. Y. Yang et al., *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp. 877-880, 1995.
[23] H. -S. Wong et al., *Int. Electron Devices Meeting (IEDM) Tech. Dig.*, pp.747-750, 1994.
[24] I. Y. Yang et al., *IEEE, Trans. on Eltrr. Dev.,* vo. 44, pp 822-831, 1997.