FinFET SRAM – Device and Circuit Design Considerations

Hari Ananthan, Aditya Bansal and Kaushik Roy

Dept. of Electrical and Computer Engineering

Purdue University, West Lafayette, IN 47905

hanantha@purdue.edu

Abstract

The quasi-planar double-gate FinFET has emerged as one of the most likely successors to the classical planar MOSFET for ultimate scalability. Unlike planar devices, its channel width is in the vertical direction; hence it is possible to increase effective channel width (and hence drive current) per unit planar area by increasing fin-height (SOI thickness). This translates directly to improved performance in interconnect-dominated circuits. In this paper, we explore the joint V_{dd}–fin-height–V_t design space for a 65nm FinFET SRAM. We report that 69% taller fins can accommodate 18% (140mV) lower V_{dd} as well as 35% (70mV) higher V_t to deliver iso-performance at 87% lower sub-threshold leakage, 50% lower gate leakage, 25% lower dynamic energy, 13% higher static noise margin and 38% higher critical charge for soft-error immunity.

1. Introduction

The intrinsic-body double-gate MOSFET has emerged as one of the leading candidates to replace Bulk and Partially-Depleted SOI CMOS due to its superior scalability for a given gate insulator thickness, better short-channel behaviour without complex channel engineering, higher mobility and the absence of random dopant fluctuation effects. The ideal MOSFET is essentially a gate-voltage controlled switch, and the short channel effect reflects the negative influence of drain-voltage on channel electrostatics as channel length decreases. The double-gate fully-depleted MOSFET diminishes the short-channel effect by bringing the gate closer to all regions of the channel, and thus improves scalability.

The quasi-planar SOI FinFET [6] and other variants have been proposed as easier manufacturable options compared to planar double-gate devices. Researchers have begun to develop design machinery for migration of microprocessor designs from PDSOI to FinFET CMOS [10]. Unlike planar single- and double-gate devices, the FinFET effec-
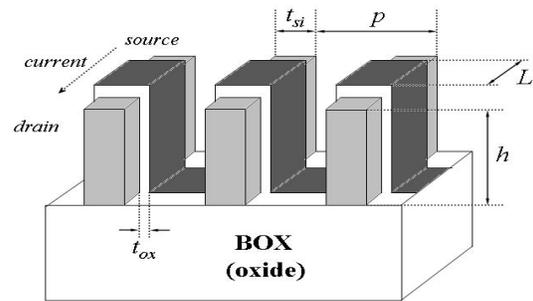

Figure 1. Multi-fin FinFET structure

tive channel width is perpendicular to the semiconductor plane. Hence it is possible to increase the effective channel width (and hence drive current) per unit planar area by increasing fin-height. Increasing drive current at the expense of gate area does not achieve performance benefits in gate-capacitance-dominated logic; delay is proportional to the ratio C_{load}/I_{drive}. On the other hand, interconnect-dominated circuits such as memory arrays are likely to benefit from the increased drive current.

An estimated 70% of the transistors in a billion-transistor superscalar microprocessor are expected to be used in memory arrays, especially large L2 and L3 SRAM data caches [12]. Thus, chip area and leakage are determined primarily by these arrays. Further, with around 3 to 5 cache accesses occurring per cycle in a 16-wide issue machine, the performance of the pipeline depends to a large extent on cache access time. The performance of an SRAM subsystem is determined primarily by the delay involved in driving large loads on the bitline and the wordline. In fully-depleted SOI, junction capacitance is negligible, so the bitline load is entirely interconnect. Hence increasing cell device widths (and hence drive current) even at the cost of higher gate capacitance decreases delay. Alternately, under a power-constrained design scenario, higher widths can

accommodate a decrease in V_{dd} and an increase in V_t (transistor threshold voltage) to save power while maintaining performance. Planar CMOS technologies (bulk or double gate) do not allow a "free" increase in channel width; the associated area penalty decreases array density and diminishes delay advantages because of the increase in wordline and bitline lengths. The quasi-planar FinFET allows an increase in effective channel width without any area penalty simply by increasing fin-height.

In this paper, we explore the joint V_{dd}–fin-height–V_t design space for a 65nm 32K FinFET SRAM array. We estimate the impact on array sub-threshold and gate leakage, dynamic energy, static noise margin and soft error immunity at iso-performance. Since all FinFETs on a die are expected to have the same height (essentially the SOI thickness), we also estimate the impact of this exploration on the performance and power of gate-capacitance-dominated logic.

2 Device Design and Simulation

Figure 1 shows the structure of a multi-fin FinFET. A silicon fin of thickness t_{si} is patterned on an SOI wafer. The gate wraps around on either side of the fin (over the gate insulator), and t_{si} is the body-thickness of the resulting double-gate structure where both gates are tied together. Current flow is parallel to the wafer plane (though occurring in an orthogonal crystal plane), while channel width is perpendicular to the plane. The effective channel width of a two-channel single-fin FinFET is thus equal to $2h$ (h = height = SOI thickness); higher widths are achieved by drawing multiple fins in parallel and wrapping the gate around them. The effective channel width for a multi-fin FinFET on a given planar area of silicon is determined by h and fin-pitch p. The fin-pitch is expected to scale as the lithography half-pitch using spacer technology [4]. The minimum h required to achieve equivalent planar area efficiency is thus $p/2$; increasing h beyond $p/2$ increases area efficiency. The upper bound on h is set by the maximum fin aspect ratio ($a_{max}=h_{max}/t_{si}$) allowed by the process. Another consideration for the upper bound is the minimum width and width-increment required in the design, since width is quantised in integer multiples of $2h$. Thus, there exists a design space for h between $p/2$ and $a_{max}t_{si}$ [17].

Figure 2 shows the two-dimensional device structure used for the symmetrical-gate FinFETs. The gate workfunctions are determined such that the 65nm logic technology ITRS node [7] on- and off-current requirements are approximately met at the nominal height ($h = p/2$). Several metals and alloys with near-mid-gap adjustable workfunctions have been demonstrated for FinFETs [3, 8]. In this work, a 70mV increase in NMOS and PMOS V_t is assumed to be achievable by adjusting the gate workfunction. In reality, this may be achieved through other means such as body
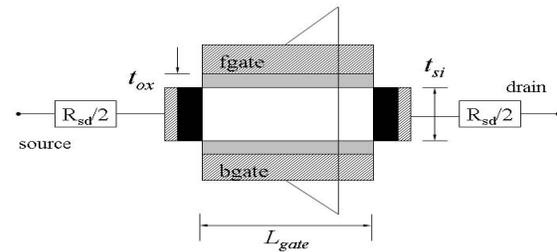
Parameter	Nominal Value
Supply Voltage (V_{dd})	0.77 V
Physical gate length (L_{gate})	25 nm
Physical oxide thickness (t_{ox})	1.0 nm
Body thickness (t_{si})	11 nm
Fin-pitch (p)	65 nm
Fin-height (h)	32.5 nm
Body Doping	intrinsic
S/D Doping	1e20 cm^{-3}
Lumped S/D Resistance (R_{sd})	140 Ω–μ
NMOS gate workfunction (wfn)	4.44 eV
I_{off}	1.58 μA/μm
I_{on}	1618 μA/μm
Sub-threshold slope	82 mV/dec
DIBL	73 mV/V
PMOS gate workfunction (wfp)	5.06 eV
I_{off}	1.37 μA/μm
I_{on}	1189 μA/μm
Sub-threshold slope	83 mV/dec
DIBL	73 mV/V

Figure 2. (a) Double-gate device simulation structure (b) Nominal 65nm device parameters

doping. Quasi-abrupt S/D junctions with no overlap are assumed; lumped resistances are used to account for S/D extension resistance.

The physical t_{ox} in this work is somewhat smaller than is required for double-gate MOSFETs of this dimension. Using a thicker oxide necessitates the use of a thinner fin to suppress the short-channel effect; this worsens the impact of process variations when fin-thickness is controlled lithographically [16]. Using a thinner fin also decreases the fin-height design space, given that the maximum aspect ratio (a_{max}) assumed is 5:1 [17]. However, researchers have reported FinFETs with higher aspect ratios [9]. We assume a small t_{ox}–large t_{si} scenario to demonstrate the benefits achievable over a large fin-height design space. It stands to reason that our observations remain valid under a larger t_{ox}–smaller t_{si} scenario; however the gains are smaller over
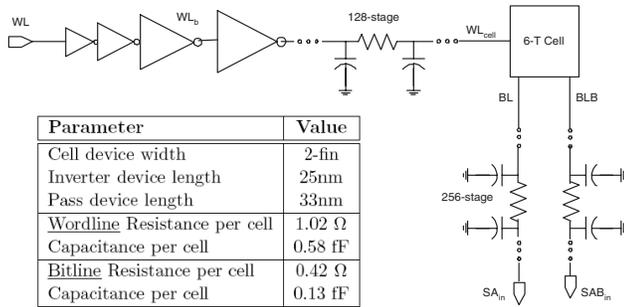
Parameter	Value
Cell device width	2-fin
Inverter device length	25nm
Pass device length	33nm
Wordline Resistance per cell	1.02 Ω
Capacitance per cell	0.58 fF
Bitline Resistance per cell	0.42 Ω
Capacitance per cell	0.13 fF

Figure 3. Circuit Model and parameters

Parameter	Design 1	Design 2	Design 3
V_{dd}	0.77V	0.7V	0.63V
h	32.5nm	42nm	55nm
wfn (array)	4.44eV	4.48eV	4.51eV
wfp (array)	5.06eV	5.02eV	4.99eV
wfn (driver)	4.44eV	4.44eV	4.44eV
wfp (driver)	5.06eV	5.06eV	5.06eV

Figure 4. Design space explored

a smaller fin-height design space.

A commercial device simulator – TAURUS [13] – is used to run two-dimensional device-circuit simulations. The Caughey-Thomas high-field mobility model is assumed for drift-diffusion transport. Quantum confinement effects are accounted for by solving one-dimensional Schroedinger equation (gate-field direction) self-consistently. Gate-oxide tunneling is solved self-consistently with majority and minority carrier transport for leakage estimation.

3 Circuit Model

Figure 3 shows the circuit model and the parameters used for the array. A 32K 6-T SRAM is organized as a 128 column–256 row array. A thin-cell layout is assumed, and interconnect RCs are adapted from previous 3-D simulations [14, 15]. Cell device dimensions are extrapolated from a previously reported FinFET SRAM structure [11]. The cell is verified to be readable and writable under worst-case ±30mV (15%) mismatch in cell device V_t's. The wordline and the bitline are modeled as distributed pi-RC networks. The pass-transistor gate capacitances are derived from C-V simulations. Junction capacitances are neglected because of the fully-depleted nature of the devices. A four-stage Fan-out-4 (last stage–Fan-out-6) wordline driver is designed with symmetrical rise and fall times. The input signal at the wordline driver (WL) is assumed to have a slew rate of 10ps.

Figure 4 shows the design space explored in this experiment. Design 1 is the starting nominal fin-height design (with larger V_{dd} and smaller V_t), and Design 3 is the final maximum fin-height design (with smaller V_{dd} and larger V_t). V_{dd} and h are varied for the cell and the wordline driver. The gate workfunctions are varied only for the cell devices; the wordline driver is maintained at the nominal (low-V_t) value. We assume that the S/D extension sheet resistance is the dominant component of R_{sd}; hence increasing h increases extension cross-sectional area and decreases R_{sd} linearly.

4 Results and Discussion

4.1 Delay

We consider two components of SRAM delay – wordline driver (wordline driver input $\rightarrow$ SRAM cell) and bitline (SRAM cell $\rightarrow$ Sense Amplifier). The signals WL, WL_b, WL_{cell}, BL and BLB are as shown in Figure 3. The delay can be expressed as –

$$
\begin{aligned}
\tau_{sram} &= \tau_{wldriver} + \tau_{bl} \\
&= (3\tau_{inv} + \tau_{wl}) + \frac{C_{bl}\Delta V_{sense}}{I_{on-cell}} \\
&\approx (3\tau_{inv} + \frac{C_{wl}V_{dd}}{I_{on-driver}}) + \frac{C_{bl}\Delta V_{sense}}{I_{on-cell}},
\end{aligned}
$$

where

$$
\begin{aligned}
3\tau_{inv} &: \quad WL \rightarrow WL_b, \\
\tau_{wl} &: \quad WL_b \rightarrow WL_{cell}, \\
\tau_{inv} &= (\frac{C_{ox}LV_{dd}}{C_{ox}(V_{dd}-V_{on})v_{sat}})(\frac{W_{load}}{W_{driver}}), \\
C_{wl} &= C_{wl-pass} + C_{wl-int}, \\
C_{bl} &= C_{bl-int}, \\
I_{on-driver} &= C_{ox}W(V_{dd}-V_{on})v_{sat}, \\
I_{on-cell} &= I_{on(M5-M1)}.
\end{aligned}
$$

τ_{bl} is the time required for a differential voltage ΔV_{sense} (50mV) to develop between BL and BLB, after which the sense amplifier gets activated. $I_{on-cell}$ is the cell pull down current through transistors $M5$ and $M1$ (from Figure 6(a)) that discharges the bitline. The wordline and bitline interconnect capacitances (C_{wl-int} and C_{bl-int}) and ΔV_{sense} are assumed to be constant for all 3 designs. The pass transistor component of wordline load ($C_{wl-pass}$) increases linearly with increase in h.

Figure 5 shows the array waveforms. Both components of $\tau_{wldriver}$ remain nearly invariant over the design space. From design 1 to 3, the increase in τ_{inv} because of the
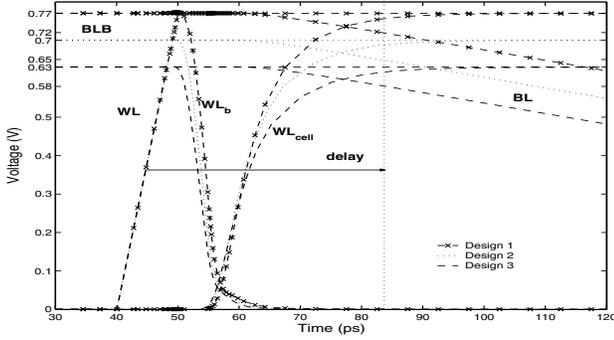
Figure 5. Simulated waveforms

increase in V_{on}/V_{dd} is very small. τ_{wl} decreases slightly – the increase in driver size (through h) and the dominance of C_{wl-int} over $C_{wl-pass}$ overrides the decrease in driver strength (through larger V_{on}/V_{dd}) and the increase in $C_{wl-pass}$ (through h). The design points were originally chosen to maintain $I_{on-cell}$ and hence the bitline discharge slope; thus τ_{bl} remains constant.

4.2 Array Leakage

Figure 6(a) shows the various leakage paths in an 6-T cell. The cell leakage power can be expressed as –

$$
\begin{aligned}
P_{leak} &= P_{sub} + P_{gate} \\
&\sim V_{dd}hI_{ds} + V_{dd}hI_g,
\end{aligned}
$$

where I_{ds} and I_g are defined per unit width.

Figure 7(a) shows the decrease in sub-threshold (-87%) and gate conduction-band-electron tunneling (-50%) cell leakage power from design 1 to 3. Increasing h increases P_{leak} linearly. Decreasing V_{dd} improves sub-threshold slope and thus decreases I_{ds}. Smaller gate field decreases I_g exponentially. Increasing V_t decreases I_{ds} exponentially. These exponential effects coupled with the decrease in V_{dd} override the impact of increasing h.

4.2.1 Gate Leakage

The gate leakage results include conduction-band-electron tunneling (CBET) for all cell devices. This accounts for the major portion of tunneling current in the NMOS devices. Valence-band-electron and valence-band-hole tunneling (VBET and VBHT) results are not available because of convergence difficulties with the simulator.

Figure 6(b) shows the various components of gate tunneling. CBET accounts for the gate-to-channel component (I_{gc}) in NMOS. VBHT is the inverse mechanism of CBET
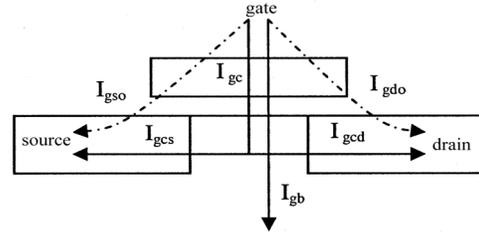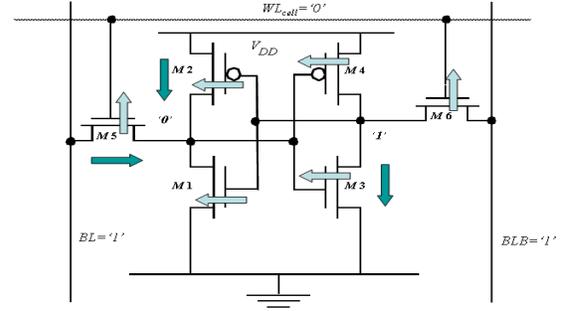

Figure 6. Components of (a) SRAM cell leakage (b) Device gate leakage

for I_{gc} in PMOS and is expected to follow a similar trend; the value of current is typically much smaller than NMOS [5]. I_{gc} is expected to be the dominant mechanism for gate tunneling in these bias regimes [2]; VBET, which accounts for gate-to-body tunneling (I_{gb}) is thus expected to be small as well. Edge-direct tunneling (EDT) from gate-to-source and gate-to-drain (I_{gso} and I_{gdo}) is dominated by CBET [1] and is thus accounted for; however, because of the absence of overlap in our devices, it does not play a significant role.
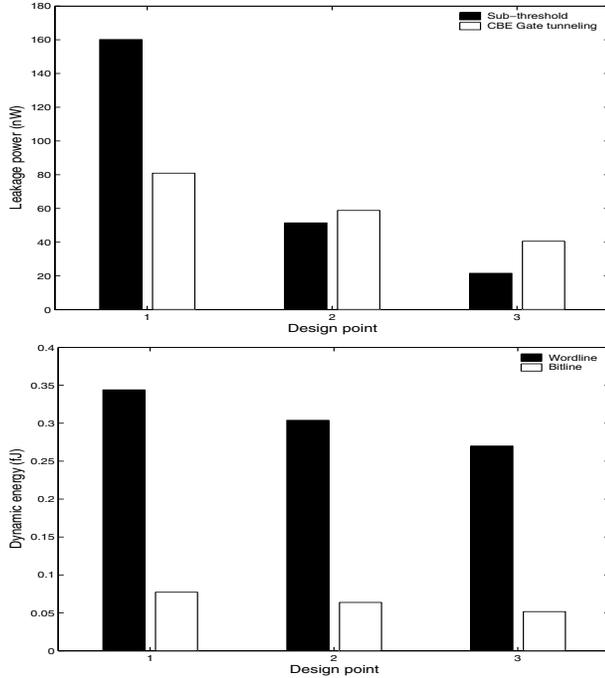
Thus, we expect that CBET is a good indication of the overall gate leakage current. Further, all components of gate tunneling have a similar exponential dependence on V_{dd} [1] and a linear dependence on h; so the overall gain is expected to follow a similar trend.

4.3 Array Dynamic Energy

Array dynamic energy is expended in charging and discharging the wordline and the bitline. The total energy during a read/write operation can be expressed as –

$$
\begin{aligned}
E_{array-dyn} &= E_{wl} + n_{word}E_{bl} \\
&\sim C_{wl}V_{dd}^2 + n_{word}C_{bl}V_{dd}^2,
\end{aligned}
$$

where n_{word} = number of bits per word, and $C_{wl} = C_{wl-pass} + C_{wl-int}$.

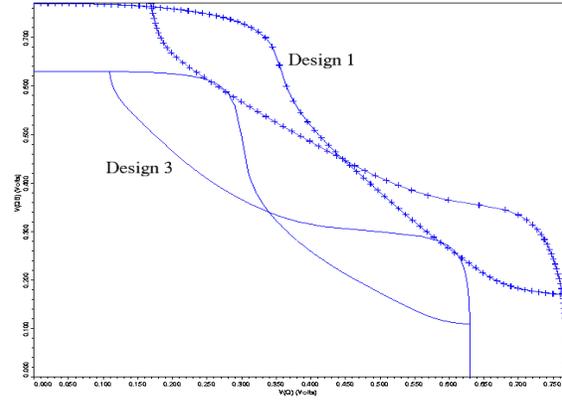**Figure 7. Results: (a) Cell leakage (b) Wordline and bitline dynamic energy**

Parameter	Design 1	Design 2	Design 3
SNM	86mV	96mV	98mV
Q_{crit} (normalised)	1.0	1.26	1.38

Figure 8. (a) Static noise margin curves (b) SNM and critical charge results

Parameter	Design 1	Design 2	Design 3
Ring-oscillator period	10.5ps	10.75ps	11.25ps
Dynamic Energy	0.194fJ	0.223fJ	0.253fJ
Sub-threshold Leakage	68nW	72nW	76nW
CBE Gate Leakage	51nW	44nW	36nW

Figure 9. Ring-oscillator Results (Dynamic energy and Leakage for a single inverter)

Figure 7(b) shows the decrease in wordline (-21%) and bitline (-33%) dynamic energy from design 1 to 3 (overall: -31%, assuming n_{word}=16). Interconnect capacitances dominate the load that is charged and discharged. A decrease in V_{dd} decreases bitline energy quadratically. The decrease in wordline energy is slightly smaller because of the increase in pass transistor component of wordline capacitance (through h).

4.4 Static Noise Margin and Soft Error Rate

Static noise margin (SNM) is defined as the side of the largest square inside the SRAM cross-coupled inverter characteristic measured during the read condition ($BL = BLB = V_{dd}$, and $WL = V_{dd}$). Figure 8(a) shows the SNM curves for the SRAM cell for designs 1 and 3. Figure 7(b) shows the increase in SNM (+13%) from design 1 to 3. Increasing V_t dominates the effect of decreasing V_{dd} and thus SNM increases.

The charge stored at the "1" node of the cell (critical charge) is usually considered a first-order indication of the extent of immunity to soft errors –

$$
\begin{aligned}
Q_{crit} &= C_{node}V_{dd} \\
&\sim hV_{dd},
\end{aligned}
$$

Figure 8(b) shows the increase in Q_{crit} (+38%) from design 1 to 3. The increase in h is greater than the decrease in V_{dd}; so Q_{crit} increases. It is not clear what effect increasing h (and hence body volume) and decreasing V_{dd} have on total charge collected during an upset event in a FinFET. Further research needs to be done to fully understand the impact on SER.

4.5 Impact on gate-capacitance-dominated logic

All FinFETs on a die are likely to have the same height (defined by SOI thickness) , and possibly the same V_{dd}. So, we estimate the impact of this design space exploration on the delay and power of a 5-stage ring oscillator. This is assumed to be representative of SRAM peripheral circuitry (decoder, wordline driver, output driver, etc.) and arithmetic units that are present on the same die as the SRAM. These devices are assumed to remain at low-V_t workfunctions.

The delay of a velocity-saturated inverter loaded by an identical gate is given by –

IEEE COMPUTER SOCIETY

$$\tau \;\; = \;\; \frac{C_{ox}WLV_{dd}}{C_{ox}W(V_{dd}-V_{on})v_{sat}}.$$

Dynamic energy and leakage power (for a single inverter) are given by –

$$
\begin{aligned}
E_{dyn} &\sim C_{ox}hLV_{dd}^2 + E_{short-circuit}, \\
P_{leak} &= V_{dd}WI_{off} \\
&\sim V_{dd}h(I_{ds}+I_g),
\end{aligned}
$$

where I_{ds} and I_g are defined per unit width.

Figure 9 shows the impact on delay (+7%), dynamic energy (+30%) and leakage power (Sub-threshold: +11%, CBE gate: -29%) from design 1 to 3. Decreasing V_{dd} increases V_{on}/V_{dd} and contributes to the increase in delay; increasing h has no effect as the driver and the load widths cancel each other. Dynamic energy increases faster than hV_{dd}^2 due to the increase in $E_{short-circuit}$ resulting from slower slew rates. Sub-threshold leakage power increases slower than hV_{dd} since I_{ds} decreases with decreasing V_{dd} due to smaller drain field. CBE gate leakage power decreases due to the exponential dependence of I_g on V_{dd}.

A larger h increases the minimum possible device width, and the quantum by which device width can be changed anywhere on the die. This might cause difficulty in designing circuits where careful balancing of widths is required, such as sense amplifiers, latches and dynamic gates [10].

5   Conclusion

The FinFET is a promising candidate for mainstream CMOS integration. The unique quasi-planar structure allows an increase in effective channel width (and hence drive current) without any area penalty by increasing device height. We exploit this property to demonstrate power savings at iso-performance in an SRAM by reducing V_{dd} and increasing V_t. In effect, we demonstrate the benefits unique to quasi-planar technologies such as FinFET (equivalent to design points 2 and 3) compared to planar bulk and double-gate technologies (equivalent to design point 1).

Similar techniques could be employed for other interconnect-dominated structures such as register files, DRAMs etc. Alternately, power-density is becoming an important issue in circuits such as ALUs and clock buffers; a smaller increase in h and/or a larger decrease in V_{dd} accompanied by a decrease in V_t could enable the designer to improve power-density while trading off leakage at iso-performance. Overall, a careful joint optimization of V_{dd}, h and V_t is required to meet system design goals.

Acknowledgements

References

[1] K. Cao et al. BSIM4 gate leakage model including source drain partition. In *IEDM Tech. Dig.*, pages 815–818, 2000.

[2] C.-H. Choi, K.-Y. Nam, Z. Yu, and R. Dutton. Impact of gate direct tunneling current on circuit performance: a simulation study. *IEEE Trans. Electron Devices*, 48(12):2823–2829, Dec 2001.

[3] Y.-K. Choi et al. FinFET process refinements for improved mobility and gate workfunction engineering. In *IEDM Tech. Dig.*, pages 259–262, 2002.

[4] Y.-K. Choi, T.-J. King, and C. Hu. Nanoscale CMOS Spacer FinFET for the terabit era. *IEEE Electron Device Lett.*, 23(1):25–27, Jan 2002.

[5] F. Hamzaoglu and M. Stan. Circuit-level techniques to control gate leakage for sub-100nm CMOS. In *Proc. Intl. Symp. Low Power Electronics and Design*, pages 60–63, 2002.

[6] D. Hisamoto et al. FinFET–a self-aligned double-gate MOSFET scalable to 20 nm. *IEEE Trans. Electron Devices*, 47(12):2320–2325, Dec 2000.

[7] International Technology Roadmap for Semiconductors 2002 Update. Semiconductor Industry Association, http://public.itrs.net.

[8] J. Kedzierski et al. Metal-gate FinFET and fully-depleted SOI devices using total gate silicidation. In *IEDM Dig.*, pages 247–250, 2002.

[9] Y. Liu, K. Ishii, T. Tsutsumi, M. Masahara, and E. Suzuki. Ideal rectangular cross-section Si-fin channel double-gate MOSFETs fabricated using orientation-dependent wet etching. *IEEE Electron Device Lett.*, 24(7):484–486, Jul 2003.

[10] T. Ludwig et al. FinFET technology for future microprocessors. In *Proc. IEEE Intl. SOI Conf.*, pages 33–34, 2003.

[11] E. Nowak et al. A functional FinFET-DGCMOS SRAM cell. In *IEDM Dig.*, pages 411–414, 2002.

[12] Y. Patt, S. Patel, M. Evers, D. Friendly, and J. Stark. One billion transistors, one uniprocessor, one chip. *IEEE Trans. Comput.*, 30(9):51–57, Sep 1997.

[13] Taurus-Device Simulator. Synopsys, 2002.

[14] K. Tomita et al. Sub-1μm^2 high density embedded SRAM technologies for 100nm generation SOC and beyond. In *Symp. VLSI Technology Dig. Tech. Papers*, pages 14–15, 2002.

[15] Y. Tsukamoto et al. Realistic scaling scenario for sub-100nm embedded SRAM based on 3-dimensional interconnect simulation. In *SISPAD*, pages 63–66, 2002.

[16] S. Xiong and J. Bokor. Sensitivity of double-gate and FinFET devices to process variations. *IEEE Trans. Electron Devices*, 50(11):2255–2261, Nov 2003.

[17] B. Yu et al. FinFET scaling to 10nm gate length. In *IEDM Dig.*, pages 251–254, 2002.