Minimum Supply Voltage for Bulk Si CMOS GSI

Azeez J. Bhavnagarwala, Blanca Austin and James D. Meindl

Microelectronics Research Center, Dept of Electrical Engr., Georgia Institute of Technology, Atlanta, GA 30332
{azeez, blanca, james.meindl}@ee.gatech.edu

1. Abstract†

Limits on energy dissipation are investigated for bulk Si CMOS circuits at each node of the 1997 National Technology Roadmap for Semiconductors (NTRS). Physical, continuous and smooth MOSFET Transregional drain current models that consider high-field effects in scaled devices, and permit trade-offs between saturation drive current and subthreshold leakage current are described and employed to model CMOS circuit performance and power dissipation at low voltages. The Transregional models are used in conjunction with physical threshold voltage roll-off models and stochastic interconnect distributions, at performances, chip sizes and transistor counts forecast by the 1997 NTRS, to project optimal supply and threshold voltages, minimizing total energy dissipated by CMOS logic circuits. Techniques exploiting datapath parallelism to further reduce supply voltage are shown to offer decreasing reductions in power dissipation with technology scaling.

2. Introduction

The 1997 NTRS [1] projects the supply voltage of future gigascale integrated systems to scale from 2.5V in 1997 to 0.5 V in 2012 to reduce power dissipation (Fig. 1), increases of which are projected to be driven by higher clock rates, higher overall capacitance and larger chip sizes. A key challenge in the design of logic circuits will be to meet the projected performances given the competing requirements of high performance and low standby power at low voltages [1,2,3] in the presence of DIBL (Drain Induced Barrier Lowering) as well as statistical variations of dopant atoms in the channel region that limit threshold voltage control for scaled devices [4]. A methodology simultaneously considering the device, circuit and system levels of the design hierarchy is employed to minimize total power dissipated from a static CMOS gate during a clock cycle [5] for each of the 1997 NTRS technology generations, given realistic operating environments of range of operating temperature, chip size, gate count, clock frequency, wiring capacitance, and critical path depth. This analysis uses physical and stochastic models, verified by HSPICE and actual microprocessor implementations to investigate opportunities to scale V_{dd} to the optimal point corresponding to the limits of CMOS power dissipation. The analysis considers both – Uniformly Doped (UD) and Retrograde Doped (RD) MOSFETs – the bulk Si alternative to a UD MOSFET that promises, in addition to higher performance, smaller feature sizes [6] (Fig 2) and higher immunity to intrinsic and extrinsic threshold voltage fluctuations as well[7]. These premises are further extended considering additional datapaths operating in parallel at reduced clock rates and supply voltages for a given throughput, measured in total number of logic operations per second and reveal that architecture-driven voltage scaling [8] will yield decreasing reductions in power dissipation with scaling due to decreasing V_{dd}/V_{to} ratios that accompany scaling of feature size along the Roadmap.

3.Circuit and Device Models

The performance of a generic CMOS processor is modeled assuming a critical path of 15, 2-way NAND stages, each stage driving average wire lengths, which are determined, in units of gate pitches, from stochastic interconnect distributions [9], derived recursively using Rent's rule, and verified for an actual microprocessor in [9]. In logic-intensive CMOS chips, packing densities are interconnect limited [10] where the effective size of a gate is determined by its wireability [11]. The gate pitch is estimated from NTRS projections for ASIC chip size and transistor count and is used in calculating the average wire length. Assuming equal interconnect cross-sectional dimensions, and that neighboring wiring planes in a multi-level network provide an approximate ground plane, total capacitance per unit length, including fringing effects, is estimated using analytical models in [12]. Circuit and device performance is modeled using compact low-voltage Transregional MOSFET models ((1)-(7) in the Appendix) that predict circuit performance in the sub-threshold, saturation and linear regions of operation (Figs. 3-5) providing continuous and smooth transitions across region boundaries [13]. High field-effects on carrier mobility are incorporated by adopting the mobility reduction model in [14]. Smoothness and continuity of the drain current expressions in the triode, saturation and the subthreshold regions are obtained by requiring differentiability and continuity of the product of the effective mobility and the areal charge density of inversion layer carriers. Low field mobility dependence on temperature and doping concentration is estimated using empirical models reported in [15]. The doping profile for the RD structure is selected as one that yields the smallest depletion depth, corresponding to the least DIBL effects for a given V_{to} and gate oxide thickness [16]. Increases in leakage current due to DIBL effects are calculated using 2D subthreshold models [6] that accurately predict the threshold voltage roll-off dependence on supply voltage, device geometries and doping profile. The 2-way NAND gate, as a basic circuit building block in the critical path, has a performance that parallels that of any other circuit actually used in processor critical paths in reflecting technology improvements [17]. The improved delay dependence on fan-in at short channel lengths [18] due to a smaller reduction in the drain saturation current with a rise in the source voltage of the topmost series-connected MOSFET is modeled physically by calculating the fractional reduction of the normalized saturation drain current for the series-connected structure (eqn 3-7 in Appendix A) Fig 4 compares this model with HSPICE simulations.

† This work was supported by the Defense Advanced Research Project Agency (Contract: F3361595C1623) and the Semiconductor Research Corporation (SJ-374-002)

4. Minimum Power CMOS

Power drain of a static CMOS gate is minimized by scaling the supply voltage while meeting the performance required by scaling the threshold voltage and increasing the channel widths until further decrease in threshold voltage increases total power due to a dominating static component [3,5,8] (Fig. 6-7). Threshold voltage decrease due to DIBL becomes less severe with a decrease in device threshold voltage due to a lower supply voltage required to meet the same performance. Further increases in channel width cause device capacitances to dominate over wiring and increase total power due to larger gates [8] (Fig. 7).

5. Parallel datapaths

Scaling the supply voltage and compensating for the performance loss by adding datapaths in parallel [8] (Figs 9&10) so that the total number of logic operations per second, or system throughput remains constant, permits the total power drain from all of the datapaths to be reduced due to a relaxed cycle time requirement on each datapath. Increasing the number of parallel datapaths, N_p increases the complexity and size of the overhead circuitry required for routing, multiplexing and control of each of the parallel datapaths resulting in the consumption of an additional component of overhead power. Latency also increases by a factor N_p over the single datapath case. This increase in overhead switching capacitance is modeled assuming a quadratic increase in overhead circuitry with number of processors (Eq (8,9) in Appendix). Decreasing V_{dd}/V_{to} ratios projected by the Roadmap increase the speed penalty for a given reduction in supply voltage, increasing the required number of parallel processors significantly to compensate for the loss in performance. Fig 10 demonstrates the decreasing reductions in total power dissipation obtainable by employing parallel datapaths.

6. Conclusions

Physical device and circuit models are employed to investigate the limits on energy dissipated during a binary logic transition for bulk Si UD and RD device structures. These permit supply voltages to scale to a minimum of 370mV for RD MOSFETs and 390mV for UD devices in the year 2012 at the performances projected by the 1997 NTRS. Optimal V_{dd}, V_{to} and gate sizes are calculated for each node of the 1997 Roadmap corresponding to minimum power. Datapath parallelism is projected to offer decreasing reductions in power dissipation with scaling.

7. References

1] The 1997 NTRS, Final Draft, SIA, Dec 1997
2] J D Meindl, Proc. IEEE, Vol. 83, No 4 Apr 1995, pg 619.
3] J Burr et al, ISSCC Dig Tech Papers, Feb 1994, pp 84-85.
4] X Tang et al, IEEE ISLPED, Aug 1996 Dig, pp 233-239
5] A Bhavnagarwala et al, IEEE ISLPED, Aug 1996 Dig, pp 193
6] B Agrawal et al, Proc. ESSDERC, Sept 1993, pp 919 – 926.
7] X Tang, Private Communication
8] A Chandrakasan et al, IEEE JSSC Vol 27, No 4, April 1992, pp 473
9] J Davis et al, Proc. 46th IEEE ECTC, pp 18.5 May 1996, pp 1002-1008
10] R W Keyes, IEEE JSSC, Vol SC-17, Dec 1982, pp 1232-1233
11] B Bakoglu, "*Circuit Interconnections and Packaging for VLSI*", Addison Wesley, 1990
12] J Chern et al, IEEE EDL Vol 13, No 1, Jan 1992, pg 32.
13] R Swanson et al,IEEE JSSC, Vol. SC-7, pp. 146-153, Apr. 1972
14] C Sodini et al, IEEE TED, Vol ED-31, No 10, October 1984, pp 1386
15] C Jacoboni et al, Solid State Electronics, No 20, Vol 77, 1977
16] B Agrawal et al, IEEE TED, Vol 43, No 2, Feb 1996, pg 365
17] G Sai Halasz, Proc. of the IEEE, Vol 83, Jan 1995, pp 20-36
18] T Sakurai et al, IEEE JSSC, Vol 28, No 1, Jan 1993, pg 40

8. Appendix :

$$(1)\ I_{subthreshold} = \frac{W}{L}\mu_o C_{ox}\left(\frac{2\frac{\eta}{\beta}}{\sqrt{1+\frac{4\eta\theta}{\beta}}+1}\right)^2 e^{\frac{\beta}{\eta}\left(V_{gs}-V_{to}-\frac{1}{2\theta}\left(\sqrt{1+\frac{4\eta\theta}{\beta}}-1\right)\right)}\left(1-e^{\frac{-\beta V_{ds}}{\eta}}\right)$$

(1)-(7): MOSFET Transregional model

$$(2)\ I_{linear} = \frac{W}{L}\left(\frac{\mu_o}{(1+\theta[V_{gs}-V_{to}])\left(1+\frac{V_{ds}}{LE_c}\right)}\right)C_{ox}\left((V_{gs}-V_{to})V_{ds}-\frac{V_{ds}^2}{2}+\frac{4}{3}\phi_F\frac{Q_{BO}}{C_{ox}}\left[\left(1+\frac{V_{ds}}{2\phi_F}\right)^{\frac{3}{2}}-\left(1+\frac{V_{ds}}{2\phi_F}\right)\right]\right)$$

$$(3)\ I_{sat} \cong \frac{W}{L}\left(\frac{\mu_o}{(1+\theta[V_{dd}-V_{to}])}\right)C_{ox}\left\{\frac{1}{\left(1+\frac{V_{dsat}}{LE_c}\right)}\left((V_{dd}-V_{to})V_{dsat}-\frac{V^2{}_{dsat}}{2}+\frac{4}{3}\phi_F\frac{Q_{BO}}{C_{ox}}\left[\left(1+\frac{V_{dsat}}{2|\phi_F|}\right)^{\frac{3}{2}}-\left(1+\frac{3V_{dsat}}{2|\phi_F|}\right)\right]\right)\right\}$$

$$(4)\ V_{dsat} = LE_c\left(\sqrt{1+\frac{2(V_{dd}-V_{to})}{LE_c\eta}}-1\right)\quad (5)\ t_{pd} = t_{pdo}\times f_{ineff}\quad (6)\ f_{ineff} = 1+\frac{2(N-1)V_{dsat}\left(1-\frac{1}{\sqrt{2}}\right)(1+\gamma)}{\left[(V_{dd}-V_{to})-\frac{V_{dsat}}{2}\right]}\quad (7)\ \gamma = \frac{1}{c_{ox}}\sqrt{\frac{qN_a\varepsilon_s}{4\phi_F}}$$

Overhead capacitance, Average power dissipation per gate normalized by number of gates in datapath

$$(8)\ \frac{C_{overhead}}{C_{datapath}} = mN_p^{\omega}+\Gamma\quad (9)\ P_{total}^{normalized} = \frac{1}{2}aC_L V_{dd}^2 F_o\left[1+\frac{C_{overhead}}{C_{datapath}}\right]+P_{static}^{gate}N_p$$

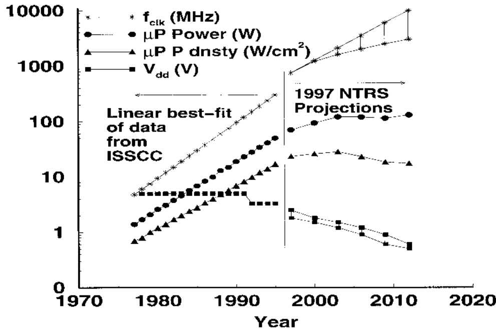


***Fig.1**: Historical trends with 1997 NTRS projections*

Year	'97	'99	'01	'03	'06	'09	'12
f_{clk}(GHz)	.75	1.2	1.4	1.6	2.0	2.5	3.0
V_{ddopt} (UD) (V)	0.94	0.84	0.75	0.71	0.58	0.47	0.39
V_{ddopt} (RD) (V)	0.81	0.75	0.65	0.61	0.52	0.43	0.37
C_w/C_L (RD)	0.33	0.3	0.36	0.38	0.42	0.47	0.51
P_{total} (RD)(µW)	4.19	4.81	4.05	2.86	2.46	1.59	1.11
P_{total} (UD)(µW)	6.20	7.75	6.8	4.95	4.32	2.76	1.85
½ $C_L V^2$(RD)(fJ)	42.5	29.6	20.9	12.9	8.3	4.3	2.5

Table-1: Optimal V_{dd}, C_w/C_L, P_{total}/gate and limits on $E_{switching}$

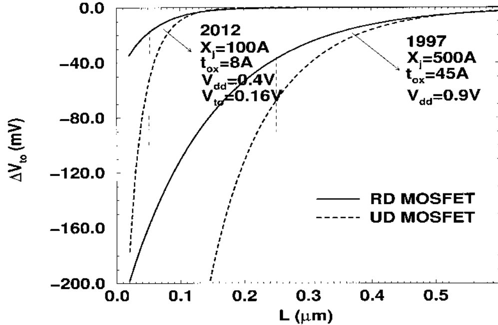


***Figure 2:** Calculated V_{to} roll-off for bulk Si at NTRS projected gate oxide thickness [6]*

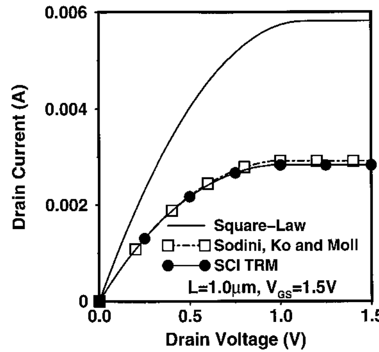


***Figure 3:** Low voltage Transregional model (TRM) for scaled MOSFETs*

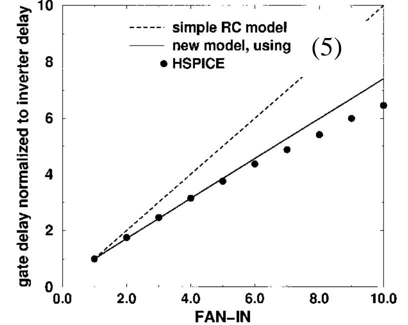


***Figure 4:** Gate delay dependence on fan-in for 0.18 micron CMOS*

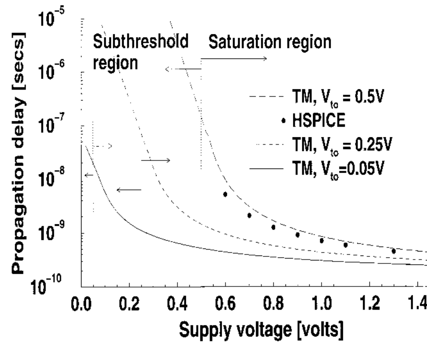
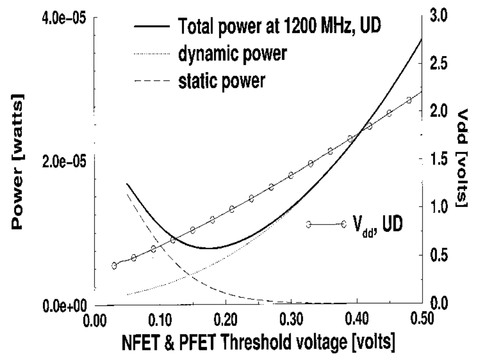
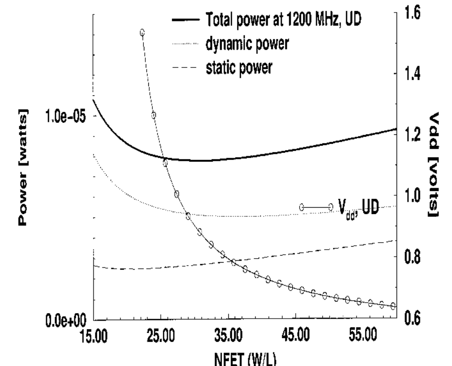


***Figure 5:** HSPICE verification of Transregional model for 0.25 µm CMOS*





***Figures 6& 7:** Power dissipation per gate dependence on V_{to}, $(W/L)_n$ and V_{dd}. L=0.18 µm*

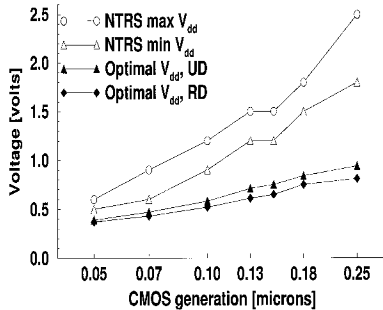
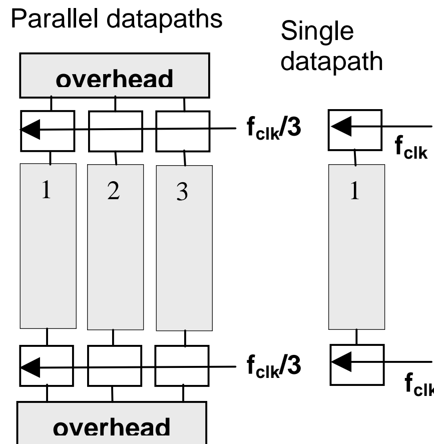
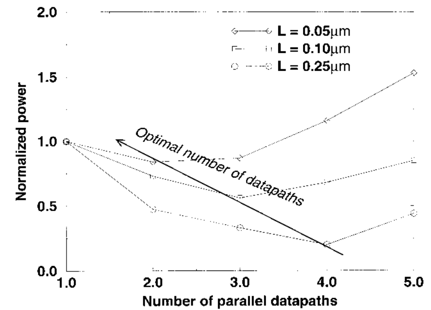


***Figure 8:** Optimal V_{dd} and 1997 NTRS projections. Optimal V_{dd} correspond to limits of CMOS power dissipation at cycle times forecasted by the 1997 NTRS*





***Figs 9 & 10:** Opportunities to scale V_{dd} using parallel datapaths [8] will yield decreasing reductions in total power with scaling of feature size.*