Power Optimization of Future Transistors and a Resulting Global Comparison Standard

P. Kapur, R. S. Shenoy, A.K. Chao, Y. Nishi and K. C. Saraswat
Center for Integrated Systems (CIS 112), Stanford University, Stanford, CA 94305
kapurp@stanford.edu

Abstract

We report a global power minimization methodology for future transistors and use it as a comparison standard to quantify the relative and absolute impact of material and structural innovations on power and speed. In addition, we put the relative tradeoffs of device design in perspective of global metrics. In the process, we also develop and verify two key enabling models: 1. to calculate inverter delay and 2. to estimate gate leakage. Although applicable to any futuristic technology node, we specifically target 45nm high-performance node with an 18nm gate length double gate FET.

Introduction

A number of devices have been proposed to replace bulk Si MOSFETs at future technology nodes (1). These novel devices result from innovations in materials (eg. high-k dielectrics, strained Si, Ge channels) and/or structure (multi-gate FETs). It is critical to be able to compare a fully optimized version of each device in the framework of standard metrics. In this work, we develop a comparison standard and exemplify its utility using extensive device simulations coupled with post processing of the simulation results. Total power and delay are the objective functions that are optimized using supply voltage V_{dd}, threshold voltage V_t, and effective oxide thickness T_{ox} as the design variables.

Methodology and Verification

We develop the framework of power-delay optimization using an 18nm gate length (L_g) double-gate FET (DGFET) (Fig. 1) as a baseline. For a given structure, we pick a delay and minimize the weighted sum of dynamic (DP), source/drain leakage (SDL) and gate leakage (GL) powers using optimal V_{dd}, V_t and T_{ox}. We repeat the above process for various delays and obtain optimum power-delay curves. The only input of this methodology is a structure specific family of curves plotting drain current, I_d and gate capacitance, C_g as a function of gate and drain voltage (V_g, V_d). These are obtained using a device simulator, Medici^{TM} (2). Although, drift-diffusion based simulations were ultimately used, a calibration comparison with hydrodynamic models reveals a 30-40% discrepancy (Fig. 2). The calculations of SDL and GL require a robust FO1 delay model, which is valid over a large range of V_t, V_{dd} and T_{ox}. Inadequacy of such an analytical model motivated a 2-step, 3-inverter numerical model (Fig.3), which accounts for temporal variations in both currents and capacitances during switching. Figs. 4 and 5 show the modeled typical inverter characteristics to be in reasonable agreement with that of a lot more computation-intensive Medici^{TM} mixed-mode simulations. The delays agree to within 70% over a wide range of V_{dd}, V_t and T_{ox} (V_{dd}=1.2V, varying V_t shown in Fig.6). The underestimation in delay is because the Miller capacitance of the subsequent inverter is not accounted. This is somewhat offset by aforementioned underestimation in I_d.

Using this delay model, we obtain SDL vs. delay curves with gate work function (ϕ_m) varying implicitly for a range of V_{dd}. (Fig. 7). ϕ_m is used to set V_t. These curves are used to obtain SDL vs. V_{dd} for fixed delay as demonstrated in Fig. 7. For calculating GL, we combine a modified model by Lee et. al. (4) with an analytical DGFET model by Taur (5). Fig. 8 shows an excellent agreement of our GL model with Medici^{TM}-simulated gate currents over 10 orders of magnitude under varying V_{dd} and T_{ox}. Finally, the switching activity (SA) specific DP is calculated assuming a clock period 16 times FO4 inverter delay. Once we have the three powers, we take their weighted sum, assuming that both '0' and '1' states of the inverter are equally likely, PMOS is twice the size of NMOS, and PMOS GL is negligible.

Results

We apply this methodology to four broad categories of problems: 1. the relative efficacy of structural (DGFET, backgate FET (BGFET), parasitic resistance) and material innovations (high-k vs. SiO_2 gate dielectric), 2. function (register vs. clock vs. datapath) specific optimal V_{dd}, 3. optimization of L_g for a given body thickness (T_{si}), 4. the impact of process and V_{dd} variations. Fig. 9 clearly exhibits a global optimum in total weighted power with respect to T_{ox} and V_{dd}. The ratio of SDL and GL to DP is approximately 20% and 5%, respectively at the optimal point. Fig. 10 explicitly quantifies the advantage of high-k dielectrics for varying delays while also revealing the optimal T_{ox} for the leaky SiO_2. High-k curves assume no degradation in mobility and short channel effects (SCE). Only a meager (~12% at highest delays and S.A.=10%) advantage with high-k is due to the robustness of DGFETs to SCE, thus yielding only an incremental improvement with T_{ox} scaling. In addition, the absence of GL in PMOS weights it less compared to dynamic power and SDL for an inverter. A lower SA yields a larger high-k advantage (Fig. 11). Fig. 12 compares the relative

0-7803-8684-1/04/$20.00 ©2004 IEEE

impact of improvement in series resistance, high-k insertion, and DGFET over BGFET in terms of global power-delay metric. A reduction in parasitic resistance results in a greater impact compared to high-k. A dramatic rise in power below a certain delay is also noteworthy. As for optimal parameters, we find that higher parasitic resistance devices need a higher optimal V_{dd}. In general, devices with better electrostatics and/or mobility/parasitics always render a lower optimal V_{dd}.

Next, we apply this methodology to obtain optimal V_{dd} for varying S.A. functions on a chip. Fig. 13 shows that the devices used in latch/clock (highest SA) need to be at lowest V_{dd}, followed by those used in datapath and finally registers at highest V_{dd}. As another application, we obtain a shallow optimum L_g for a fixed DGFET T_{si} ((2-3)T_{si}, Fig.14). The optimum for a given delay arises since a large L_g requires a lower V_t, but has better SCE while too low an L_g tolerates a higher V_t but yields poor SCE. Fig. 15 plots the impact of lowering T_{si} on optimal power-delay curves. Two trends standout: Firstly, lowering T_{si} renders a lower power despite increase in series resistance, indicating that in the considered range, improvement in SCE dominates. Secondly, the optimal T_{ox} (1.4 to 1.8nm) increases with lower T_{si} because the electrostatics are already so good that decreasing T_{ox} hurts dynamic power more than improving SCE.

Finally, the impact of L_g and V_{dd} variation on the power optimization curves is shown in Fig. 16. We observe a large power penalty with small variation along with higher optimal V_{dd}. The variations are incorporated by tuning ϕ_m such that even under the worst case (low V_{dd} and high L_g), the delay is met in the critical path. Using this ϕ_m, the worst-case static power (high V_{dd} and low L_g) is calculated and the entire optimization is repeated. Fig. 14 further shows that process variations shift the optimal L_g to higher values with an accompanying power penalty.

Conclusions

We develop an optimized power-delay comparison standard for future transistors and apply it toward various applications. 1) We quantify the advantages of DGFETs, insertion of high-k gate dielectrics, and reduction of parasitic resistance in DGFETs. We find the reduction of parasitic resistance to be more effective compared to insertion of high-k. 2) We show the optimal V_{dd}, V_t and T_{ox} values for various structures and delays. 3) With the aim of minimizing power using multiple V_{dd} and various flavors of transistors, we find that low SA circuits such as registers benefit more from high-K dielectrics than clocks and data path. In addition, low SA transistors require higher V_{dd} for power minimization. 4) For DGFETs, we report a delay-sensitive optimum gate length, which minimizes the total power. In addition, we show the need to reverse scale T_{ox}, since the minimum total power occurs at a higher EOT with thinner bodies. 5) Finally, we report the impact of V_{dd} and process variations on the optimal operating parameters and quantify the resulting power penalty. We use above conclusions only as a vehicle to elucidate the versatility of the methodology, implicitly pointing to the extensiveness and importance of its scope.

References

(1) International Technology Roadmap for Semiconductors, SIA, 2003.
(2) Medici version 2003.12, Synopsys, CA
(3) M. H. Na, E. J. Nowak, W. Haensch, and J. Cai, "The effective drive current in CMOS inverters", *IEDM Tech. Dig.*, pp. 121-124, 2002.
(4) W-C. Lee and C. Hu, "Modeling CMOS tunneling currents through ultrathin gate oxide due to conduction- and valence-band electron and hole tunneling", *IEEE Transactions on Electron Devices,* vol .48, no.7, pp. 1366-73, 2001.
(5) Y. Taur, "An analytical solution to a double-gate MOSFET with undoped body", *IEEE Electron Device Letters*, vol. 21, no.5, pp. 245-47, 2000.

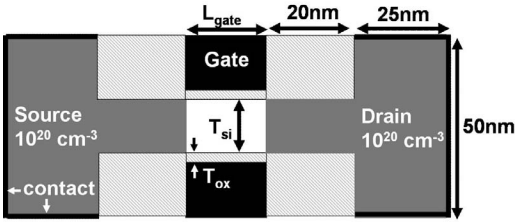

Fig. 1 Schematic of Medici-simulated double-gate (DG) FET based on ITRS 2003 45nm HP/32nm LSTP node. Baseline DG: L_g=18nm, T_{si}=7nm. BG FET:grounded bottom gate with midgap workfunction.

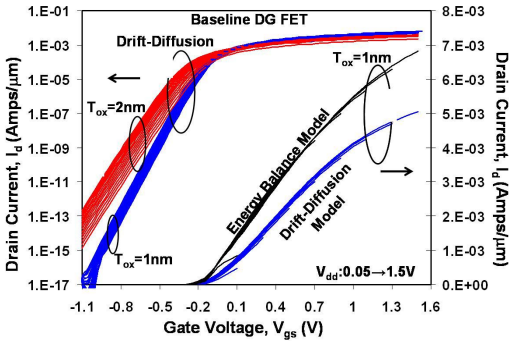

Fig. 2 Family of I_d-V_g curves with varying V_{dd} for 2 values of T_{ox}. Negative V_g is used to get effective gate workfunction Φ_m, which takes on all values within the Si bandgap. Right side axis shows 30-40% underestimate of drive current using drift-diffusion model compared to energy balance (hydrodynamic) model.

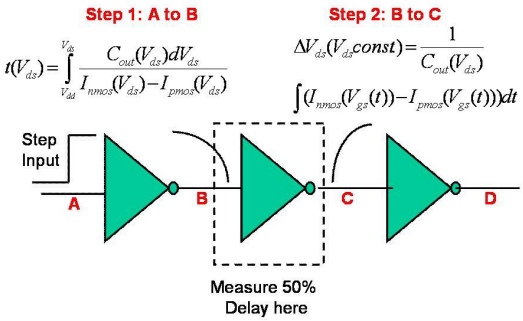

Fig. 3 Setup for FO1 inverter delay calculation by numerical integration of Medici I_d-V_g and C_g-V_g.

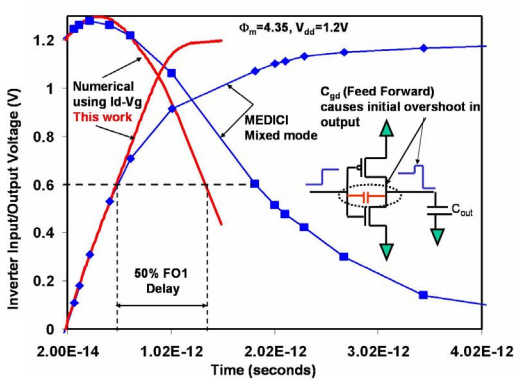

Fig. 4 Comparison of inverter delay calculation with Medici mixed-mode transient simulation.

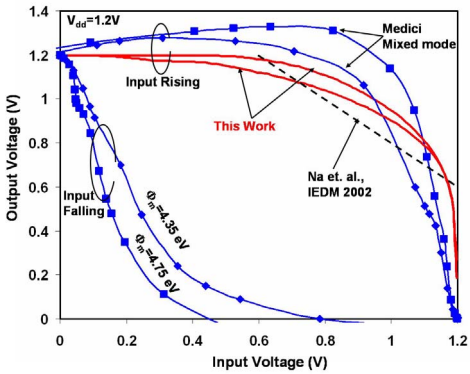

Fig. 5 Inverter switching trajectory and model comparison with Medici mixed mode transient as well as previous IBM model (3).

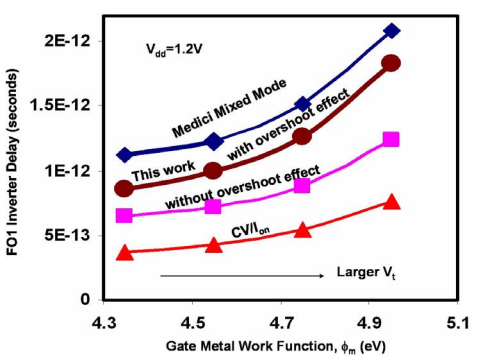

Fig. 6 Comparison of inverter FO1 delay calculation methods. Simple CV/I_{on} severely underestimates delay.

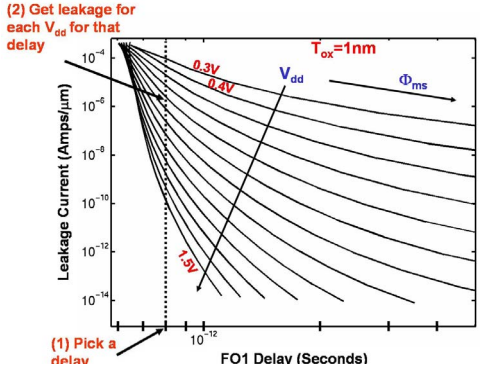

Fig. 7 Source-drain leakage power calculation. At a given delay, as V_{dd} is swept, implicit V_t adjustment sets the leakage current.

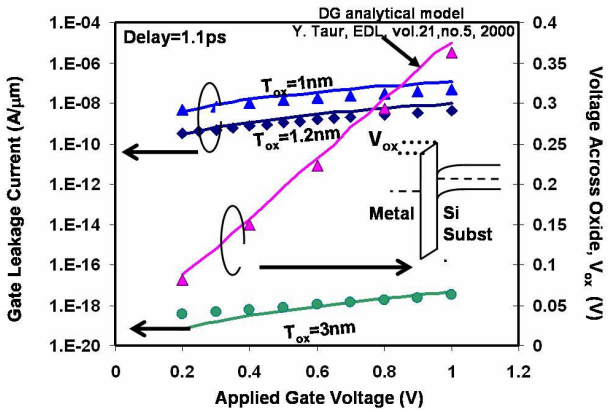

Fig. 8 Gate current and oxide voltage drop model (lines) and Medici simulation (symbols).

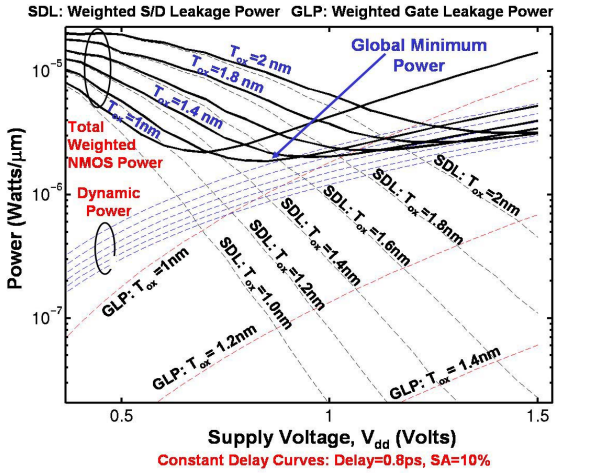

Fig. 9 Total power optimization curve for a target delay and switching activity. Dashed curves show weighted individual components of power – dynamic, S/D leakage (SDL), and gate leakage (GLP).

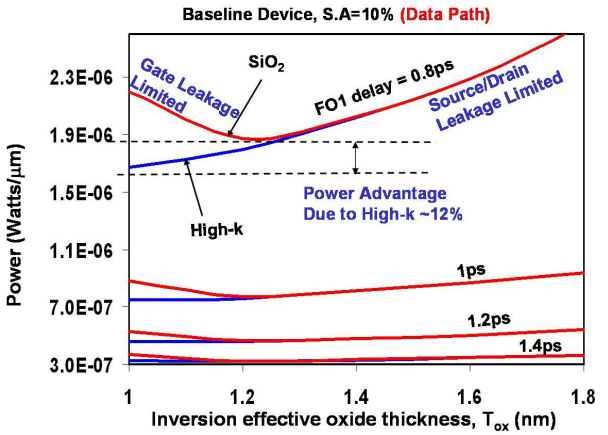

Fig. 10 Optimized total power as a function of effective inversion T_{ox} comparing SiO_2 with high-k for device with moderate switching activity (SA) of 10%.

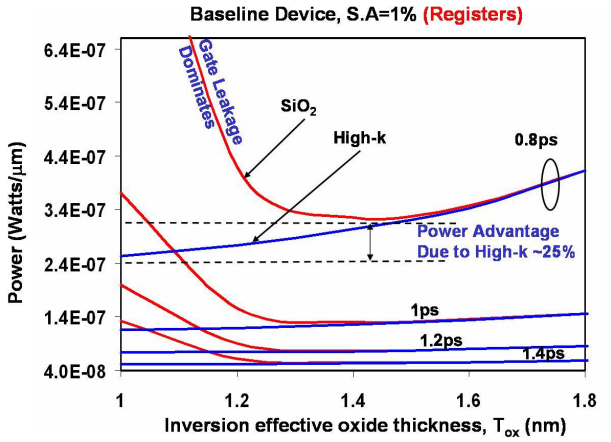

Fig. 11 Same as Fig. 10, but for device with low S.A. (1%).

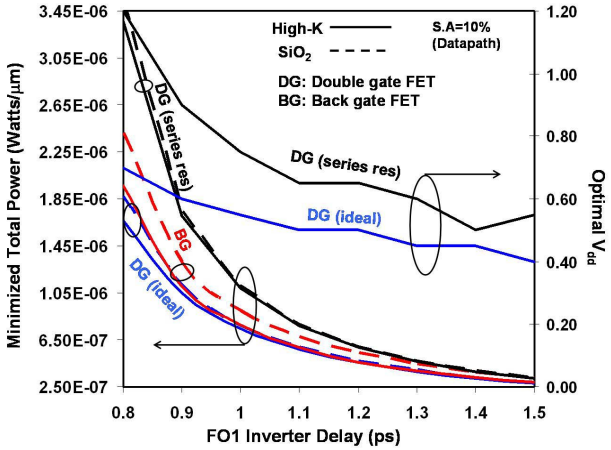

Fig. 12 Comparison of devices using optimal power-delay curves. Impact of material change (high-k vs. SiO_2), structure change (DG vs. BG), and series resistance is shown. Optimal V_{dd} for the DG with and without series resistance is plotted on the right side axis.

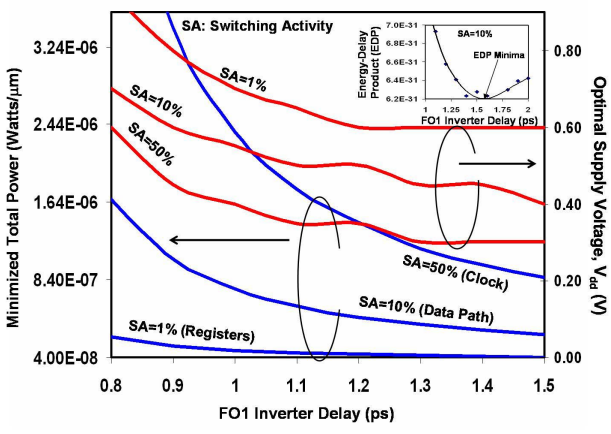

Fig. 13 Minimum total power-delay curves with corresponding optimal V_{dd} for devices with different switching activity. Inset graph shows minimization of the energy-delay product (EDP).

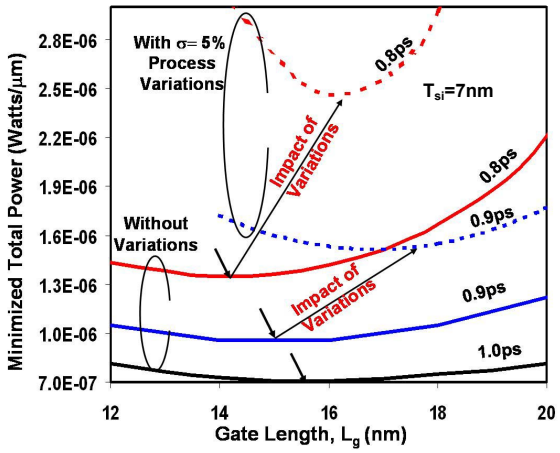

Fig. 14 Optimal gate length, L_g that minimizes total power for a given target delay and structure (baseline DG with T_{si}=7nm). Incorporating process variations increases the optimal L_g.

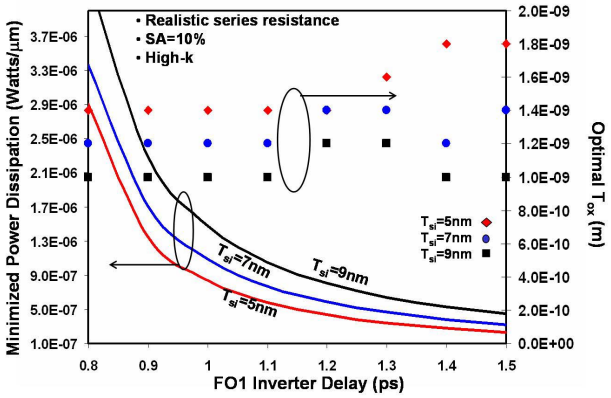

Fig. 15 Minimum total power-delay curves, along with corresponding optimal T_{ox} for DG with varying body thickness (T_{si}).

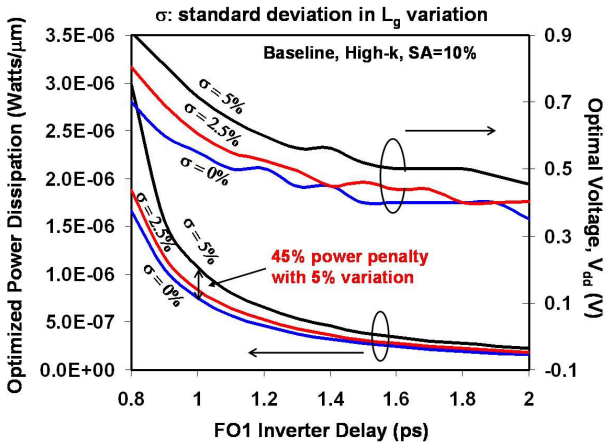

Fig. 16 Impact of process-induced gate length variations on minimum power-delay curves and corresponding optimal V_{dd}.