RC22379 (W0201-039) January 4, 2002
Electrical Engineering

IBM Research Report

Overscaling - Design for the Future

Paul M. Solomon
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Ihsan Djomehri
MIT
Department of Electrical Engineering
Cambridge, MA

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Overscaling- Design for the Future

Paul M. Solomon, and Ihsan J. Djomehri*

IBM T.J. Watson Research Ctr., Yorktown Heights, NY, *MIT, Dept. Electrical Eng., Cambridge, MA.

Abstract

A new design approach is presented for FETs at their limit of scaling where the power-performance trade-off is achieved by intentional channel length variation. The method was applied via simulation to double gate and thin SOI FETs and it was shown that a single, midgap metal was able to handle the entire power-performance range in an almost optimal manner, obviating the need for multiple work-function gates. For bulk N-FETs a single work function of ~0.1V below midgap was near optimal.

Introduction

As CMOS technology continues to evolve, with limits to CMOS scaling fast approaching, straightforward design procedures of the past, where standby power was never a serious constraint, are being replaced by multi-threshold designs where higher performance is traded for higher standby power on a selected subset of transistors. On the other hand innovative devices such as the double-gate FET (DGFET) and the ultra-thin SOI single gate FET (SGFET) are being explored for their superior performance and scaling potential, and freedom from the problem of dopant fluctuations. However; the threshold voltage of these FETs is not easily adjusted. Multiple metal work functions are proposed for this purpose, but this is difficult to implement in practise.

A method of threshold voltage control that 'comes for free' is the change of threshold voltage, V_T, with channel length, L (the 'V_T rolloff' curve). We call this 'overscaling' because L is reduced well into the forbidden V_T rolloff region. This method has been shunned in the past because V_T is difficult to control in this region, and the sub-threshold slope degrades. While these objections are qualitatively sound, they bear quantitative examination. Rather than consider V_T, which is a rather poorly defined term in this region, it is better to consider the standby current directly.

In the DGFET etc. the performance vs. standby power trade-off may be achieved by changing L (overscaling), changing the workfunction, (ϕ, referenced to mid-gap) or both. Obviously, by changing both an optimal solution can be obtained, but, in this work we shall examine how almost optimal results can be obtained by varying L alone.

Method

To illustrate our method we analyze the DGFET shown in Fig. 1, which we simulate using MEDICI™. The vertical dimensions are close to the scaling limit: 0.825nm for the oxide thickness and 5nm for the silicon thickness. The source/drain (S/D) doping falls-off at 1nm/dec, with the channel length being defined by the $2x10^{19}$ cm^{-3} doping contour. For SGFET simulations the bottom gate in Fig. 1 is simply replaced by oxide. A super-halo [1] bulk device was also simulated.

L is varied, and in addition a statistical variability, δL, is introduced. We assume δL is constant for a given technology, and the statistics are Gaussian where the magnitude of δL is at 3σ.

Drain current vs. gate voltage curves are shown in Fig. 2 for different L. According to conventional criteria, the nominal channel length would be ~16nm for the DGFET and ~20nm for the SGFET and bulk cases. We investigate the overscaling regime all the way down to 6.5nm (Over this range, we determined, using the method of Likharev [2], that S/D tunneling was negligible compared to thermionic emission current.).

Standby power (at V_G=0) vs. L is shown in Fig. 3. The various ϕ are simply obtained by shifting the gate voltage axis by the required amount. Note that unlike the steep V_T rolloff curve [3], standby power shows no 'cliff' at short channel lengths. Our *power metric*, P_{stby}, is the standby power averaged over the gate length distribution. This is *not* the nominal power, as illustrated in the inset of Fig. 3.

For the *performance metric*, F, we compute the charge control inverse time constant given by $(I_{D,on}\text{-}I_{D,off})/(Q_{on}\text{-}Q_{off})$, where charge Q_{on} and Q_{off} are measured in the low (V_D low, V_G high) vs. high logic states. Results are shown in Fig. 4 with the different work functions obtained as before. Gate length variability is accounted for by computing the worst case performance defined as F at $L+\delta L$.

Results

Worst case performance vs. standby power plots for the DGFET and are shown in Fig. 5 for σ=2nm.The plots have a sharp upper envelope, known as a 'caustic' where optimal solutions are obtained. Note that the individual curves hug this caustic, and that the curve for (ϕ=0) hugs it over most of the useful power range (from 1mW/m to 100W/m). A similar exercise for the SGFET results in Fig. 5b, where this behavior is even more striking. This is because the more gradual V_T rolloff curve of the SGFET is more favorable to overscaling the opposite is the case for the bulk superhalo FET (Fig. 5c) yet even here a useful range is obtained, with ϕ =-0.1V.[1] Fig. 6 compares the performance for δL=±1nm vs. ±3nm.The extra variability reduces the performance by ~20% but actually reinforces this overscaling argument in that a single ϕ now covers the whole power range.

In Fig. 5 results are also shown at constant L_{nom} but varying ϕ. The power-performance trade-off is clearly worse than that achieved by overscaling.

There are penalties associated with overscaling such as increased delay skews and increased sensitivity to thickness. Thus the overscaling method might be more applicable to planar DGFET and SGFET configurations where the SOI thickness can be tightly controlled.The sensitivities increase with ϕ, somewhat offsetting the advantage of using ϕ=0 or larger for the whole power range. An alternative strategy would be to use two work functions above and below midgap by ~0.15V, for both N and P-FETs.

The overscaling method enables one to compare different technologies under close to optimal conditions. Fig. 7 compares performance envelopes for δL=2nm and at V_{DD}=1V and 0.6V. The DGFET has a decided power-performance advantage in the low, and important mid-range applications, while the performances converge at very large, and less useful, standby powers.

Conclusions

Overscaling is shown to be a surprisingly robust means of trading standby power vs. performance for different device geometries and in the face of statistical gate length variability. It can greatly simplify the DGFET technology where a single, mid-gap work function gate material may be used.

1. It is interesting that the optimal ϕ for bulk is near mid-gap and not near the conduction band, however; the bulk case is complicated by the fact that the channel doping profile can also be varied, which was not done, apart from halo separation, in this study.

References

[1] Y. Taur, C.J. Wann and D.J. Frank, 1998 IEDM, Digest p.789.
[2] Y. Naveh and K.K. Likharev, IEEE Elec. Dev. Lett., April, 2000.
[3] H-S Wong, D. Frank and P. Solomon, 1998 IEDM, Digest p.407.

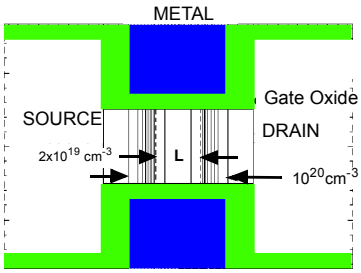

Fig. 1. Schematic cross-section of DGFET having an 0.83 nm equivalent oxide thickness and a 5nm body thickness.

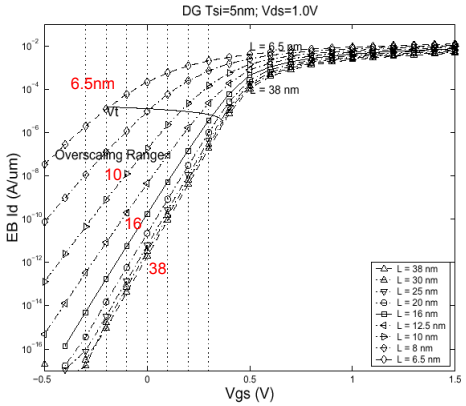

Fig. 2 Simulated drain current of the DGFET at V_{DS}=1V and for different channel lengths, showing the overscaling range.

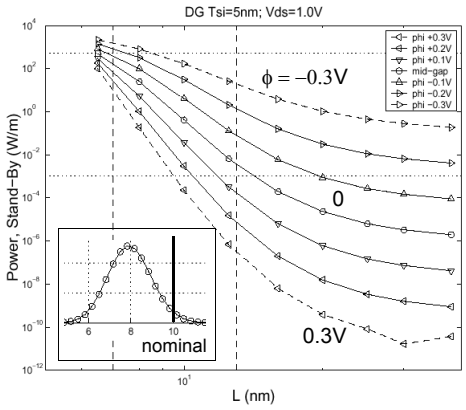

Fig. 3. Stand-by power for different gate work functions (referenced to mid-gap). Inset shows power probability density function, referenced to nominal power.

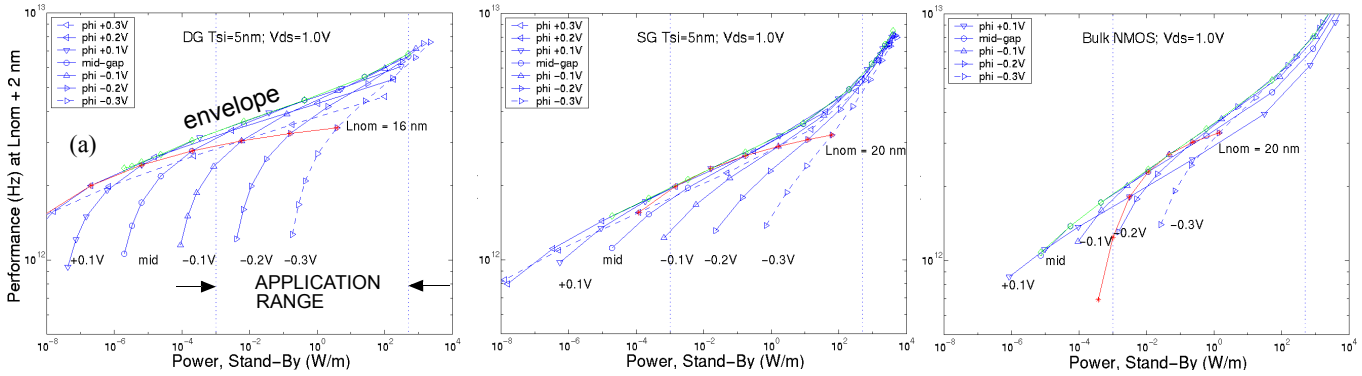

Fig. 5. Family of standby-power-performance curves (a) for a DGFET and (b) for a thin SOI SGFET and (c) for bulk FET, showing caustic envelope of optima. Dotted vertical lines indicate range of useful applications. Dashed lines with + symbols correspond to the single nominal channel length indicated.

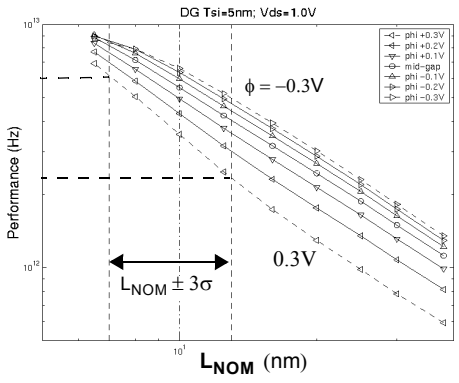

Fig. 4. Performance measure as a function of nominal channel length for various gate work-functions. Dashed vertical lines indicate channel length variability for 3σ = ± 3nm, and dashed horizontal lines indicate corresponding delay variation.

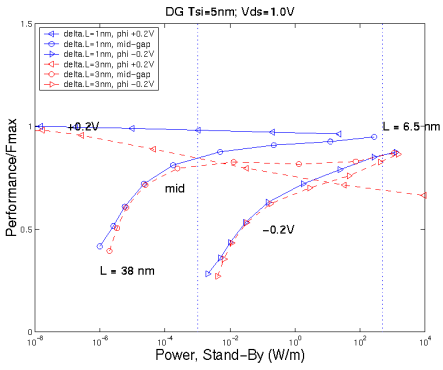

Fig. 6. Performance relative to maximum (optimal work function, zero gate-length tolerance) vs. stand-by power for different work functions and for ±1 nm (solid) and ±3 nm (3σ) linewidth variation, for a DGFET.

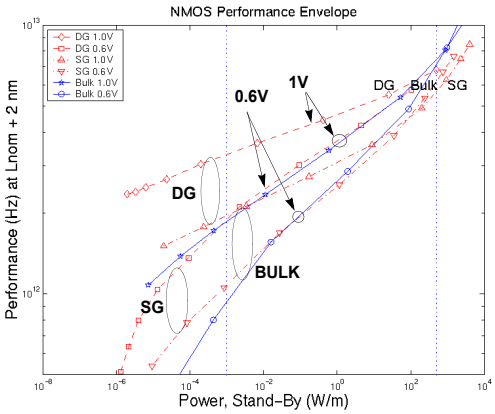

Fig. 7. Performance envelopes for DGFET, thin SOI SGFET, and bulk at V_{DD}=1.0V and 0.6V with a 3σ gate length variation of ±2nm.