G-2-1 (Invited)

Performance of Deeply-Scaled, Power-Constrained Circuits

Borivoje Nikolić, Leland Chang, Tsu-Jae King

Department of Electrical Engineering and Computer Sciences, University of California
Berkeley, CA 94720-1770, USA
Phone: +1-510-643-9297   E-mail: bora@eecs.berkeley.edu

1. Introduction

Power has become a primary design constraint in digital integrated circuits. Most designs in sub-100nm technologies will either maximize the performance under power constraints or minimize the energy for required amount of computation. To achieve the optimality in power-performance space, integrated circuits have to be optimized at all levels of hierarchy: device, circuit, microarchitecture and system architecture. The system power and performance requirements have to be propagated from the system specification all the way to the technology. In order to make optimal tradeoffs at one level of the design hierarchy, the designer must know the power-performance dependencies from the lower level [1].

At the system level, for example, performance can be traded off for power and area (cost) through adding functional units or increasing the parallelism at the system level. At the microarchitecture level, this tradeoff between the power and throughput/latency exists in the choice of parallelism level or pipelining depth. Logic designers can optimize the delay of a circuit block by optimizing its structure: for example a carry lookahead adder is faster than the ripple carry adder, but consumes more power. At the circuit level, delay and power can be traded off through sizing and the choice of supply and threshold voltages. These tradeoffs propagate all the way to the device level, where the devices can be optimized through the choice of transistor thresholds, oxide thickness, doping concentrations and profiles.

2. Scaling trends

Microprocessors have demonstrated very large improvements in performance over the past 15 years, but at the expense of increased power. While the delay was decreasing by 30% through technology scaling in each generation, the reduction in logic depths through microarchitecture changes, and slower supply scaling have resulted in doubling lead microprocessor frequencies in each technology generation [2]. Compounded with the increase in die size, this resulted in almost tripling of the power in each generation, which brought us to power densities of 100W/cm^2 today. Because of heat removal and power delivery constraints, the power will be increasing at much slower rate in future, and it is projected that it will only double over next 10 years [3].

3. Performance of Scaled Devices

Deeply scaled devices, as outlined in the roadmap [3], would allow increase in switching speeds. While the 14nm bulk-Si devices have been demonstrated [4], alternatives to planar bulk-Si have been proposed.

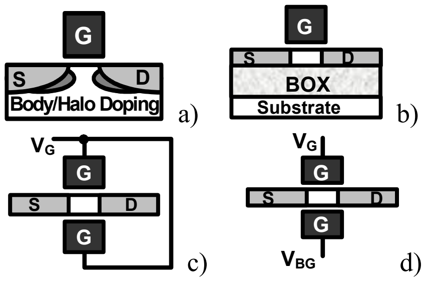

Fig. 1: a) Bulk-Si, b) Ultra-Thin-Body (UTB), c) Double-Gate (DG), and d) Ground-Plane (GP) MOSFET structure cross-sections.

Double-gate (DG) and ultra-thin body (UTB) MOSFETs (Fig. 1) have been touted as potential successors to the classical bulk-Si MOSFET [3]. Short-channel effects are effectively controlled by using a thin silicon film, allowing for gate-length scaling down to the 10nm regime [5]. In order to scale bulk-Si transistors, heavy halo doping is necessary, which degrades mobility due to impurity scattering and increased transverse electric field, increases sub-threshold slope, enhances band-to-band tunneling leakage, and increases depletion capacitance. Because thin-body devices do not require heavy channel doping, significant performance enhancements are expected [6].

To evaluate the benefits of thin-body MOSFETs from a circuit perspective, simulations are set up using realistic device structures based on ITRS specifications [3] for sub-50nm L_{gate} technology generations. Body thickness (T_{body}) requirements for a given L_{eff} are derived from scaling rules presented in [7] for DG devices; single-gate UTB devices require half this value. The minimum acceptable T_{body} may be limited to 5nm [8]. Both this case and that of unlimited T_{body} scaling are considered. Mixed-mode device simulation [9] is employed using the energy balance model for carrier transport. Because the full Boltzmann equation is not solved, drain current values may be overestimated, but the trends and differences between technologies should be valid.

The increase in I_{dsat} leads directly to an improvement in inverter delay (Fig. 2). Additional speedup (~5-10%) in thin-body devices results from the elimination of depletion and junction capacitances. Improvements over bulk devices can be as large as 45% in the DG case. This value stays relatively constant with technology scaling because the I_{off} specification increases dramatically in compliance with bulk-Si MOSFET scaling. Again, the UTB device shows a smaller enhancement, which may disappear at small gate lengths when T_{body} is limited to 5nm. The amount of improvement shown here is smaller than that reported in [10] due primarily to the realistic doping profiles used.

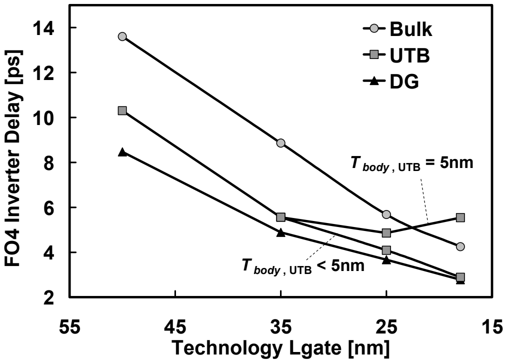

Fig. 2: FO4 inverter delay for bulk-Si, UTB and DG devices.

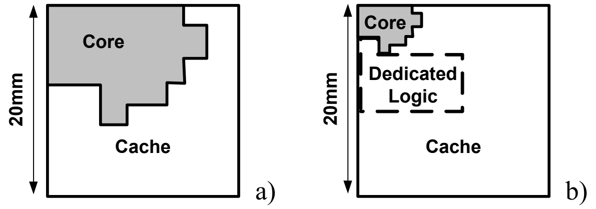

Fig. 3: Proportion of cache in microprocessor die: a) 130nm node, b) 45nm node.

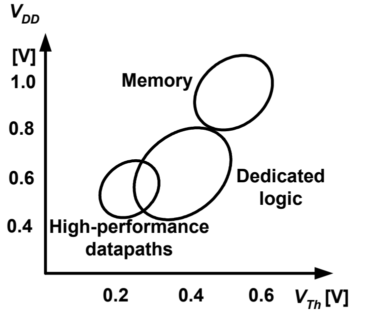

Fig. 4: Supply and threshold voltage ranges for high-performance datapaths, dedicated logic and memory.

Given a gate delay constraint, thin-body MOSFETs can also improve power dissipation by reducing V_{DD} to match the delay of a bulk-Si device. In this scenario, thin-body devices show up to a 60% reduction in energy consumption.

4. Impact of Architecture on Device Design

To achieve optimal performance in power-limited designs the design of the devices and their use in circuits should be optimized for their target application. To accommodate a variety of design targets in a single chip, multiple devices would be used. Alternatively, a single ground plane device employing back biasing could be used.

If today's microprocessor with logic depth of 14FO4 is designed in 45nm bulk-Si with L_{gate} = 18nm, it could achieve operating frequencies of over 20GHz. However, the total power density of these devices, assuming 15% activity and including leakage, would exceed 1kW/cm^2. The power density of high-performance 45nm DG and UTB devices running at 30GHz is also prohibitive.

If the lead microprocessor power is limited to about 200W, it would allow for use only of a very small percentage (<5%) of the fastest devices on the chip. Since it is difficult to increase the amount of instruction-level parallelism, it is likely that the core in high performance processors will reduce from today's 40-60% to occupy less than 10% (Fig. 3), which supports this scenario. Furthermore, in order to maintain the power density, with projected supply voltage trends, it is likely that the microarchitectures will be retargeted from today's logic depths of 14 FO4 to about 18-20 FO4 delays [11], which translates to 8-12GHz operating frequencies in bulk-Si. The increase in performance will be coming from increased amounts of on-die cache, and addition of dedicated processing units. Dedicated signal processing blocks, such as graphics processors, MPEG decoders, or networking support, employ higher levels of parallelism with longer logic depths (~50 FO4), to operate at lower frequency than the core achieving the required perceived performance.

The devices should be optimized for target logic depths, operating frequencies and activities. Highly active and fast cores would use highly leaky devices (V_{Th} ~ 0.15V in bulk Si) with approximately 0.6V supply, and higher second threshold in non-critical paths. With the total power minimized, the leakage power would present about a half of active power [1, 12]. On the other end of the device spectrum, (Fig. 4) the cache has much lower activity, which results in lower power density. To minimize the cache leakage, the devices will be using high thresholds (0.5V) with sufficiently high supplies (1V).

Dedicated datapaths, similarly to high-volume ASICs are designed with longer logic depths, and would be optimized to operate with low gate overdrive. To limit the leakage these devices would have to use aggressive leakage control techniques: e.g. in bulk-Si using power supply gating in bulk, or the ground plane in UTB.

5. Conclusions

Maximum performance in power-limited scaling regime dictates the use of variety of devices, optimized for their intended application. Double-gate and ultra-thin body devices offer improvements as large as 40% in delay and 60% in power over bulk-Si. Wide range of threshold voltages adjustments in DG devices could be used to adjust it to desired performance. Since some operating modes require low gate overdrives, these circuits will have to minimize the sensitivity to process and environment variations.

Acknowledgements

This work was supported in part by MARCO C2S2 focus center under contract 2001-CT-888.

References

[1] R. W. Brodersen, et al, *Proc ICCAD* (2002), p. 35
[2] S. Borkar, IEEE Micro 19 (1999) p. 23
[3] International Technology Roadmap for Semiconductors (2001)
[4] A. Hokazono, *IEDM* (2002), p. 639
[5] L. Chang et al., *IEDM* (2000) p. 719
[6] D. J. Frank, IBM Res J. Dev, 46, (2002), p. 235
[7] J. Kedzierski et al., *IEDM* (2001) p. 437
[8] D. J. Frank et al., *IEDM* (1992) p. 553
[9] MEDICI v2000.2 User's Manual, Avant! Corp. (2000)
[10] S. Tang et al., *ISSCC Dig. Tech Papers*, (2000) p. 118
[11] V. Srinivasan, et al, *Proc. IEEE/ACM MICRO* (2002), p. 333
[12] Gonzalez, et al, IEEE J. Solid-State Circuits 32 (1996) 1210