Moore's Law Past 32nm: The Challenges in Physics and Technology Scaling

Kelin J. Kuhn

Logic Technology Development, Intel Corporation, Hillsboro, OR, 97124, U.S.A.
Contact: kelin.ptd.kuhn@intel.com

1. Introduction

The global need for high performance and low power computing continues to be a major driver of the semiconductor industry. In the high performance computing segment, complex projects (such as medical imaging, genomics research and weather prediction) need significant performance increases to fulfill growing expectations. The core computing segment requires performance increases due to expansion into usage models which exploit life-like fully-integrated computing (such as language processing and immersive user experiences).   The small computing segment is demanding more performance at lower power and cost, with the key drivers being anytime-anywhere context-aware personalized computing.

For the past 40 years, relentless focus on Moore's Law transistor scaling has provided ever-increasing transistor performance and density. A decade ago, Moore's Law transistor scaling meant "classic" Dennard scaling [1] where oxide thickness (T_{ox}), transistor length (L_g) and transistor width (W) were scaled by a constant factor ($1/k$) in order to provide a delay improvement of $1/k$ at constant power density. However, "classic" Dennard scaling has become less influential after the 130nm node. In subsequent generations (90nm, 65nm, 45nm, 32nm, etc.) performance enhancers using new materials were added to continue to drive the transistor roadmap forward (e-SiGe, strained SiN for strain in the 90nm and 65nm nodes [2,3], and high-k metal-gate (HiK-MG) in the 45nm and 32nm nodes [4,5]). Modern CMOS scaling is increasingly a story of materials innovation.

As we look beyond 32nm planar transistors, there are a number of challenges to be addressed (Fig. 1).
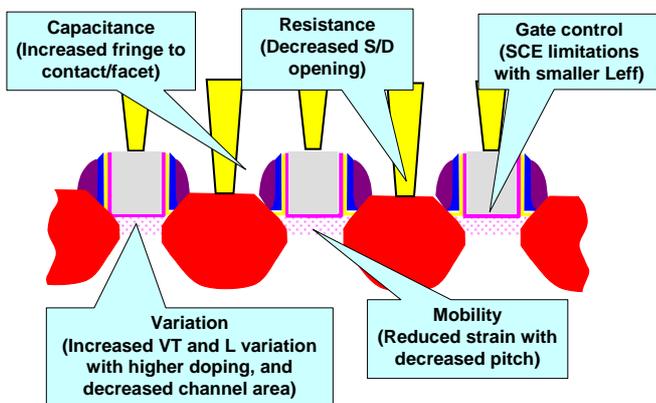

Fig. 1. Scaling challenges in MOSFET devices

Capacitance: Decreasing gate pitch increases the parasitic capacitance contribution for both contact-to-gate and epi-to-gate; thus increasing overall gate capacitance (C_{gs}). *Resistance*: Decreasing source/drain opening size increases the source drain resistance (R_{sd}); thus decreasing drive current.   *Gate control:* Decreasing T_{ox} to provide better channel control comes with a penalty of increased gate leakage current (I_{gate}) and increased channel doping. Decreasing the effective gate length (L_{eff}) without decreasing T_{ox} results in poorer short channel effects (SCE), with the consequence of increased off-state current (I_{off}) from degraded drain-induced barrier lowering (DIBL) and subthreshold slope (SS). *Mobility:* Decreasing gate pitch decreases the volume/quantity of the stressor materials for both NMOS (stress induced by overlayer films) and PMOS (stress induced by embedded-SiGe, e-SiGe); thus decreasing mobility and drive current. *Variation:* Increased channel doping combined with decreased channel area increases random V_T and L variation; thus degrading the minimum operating voltage, V_{min}.

2. Next Generation Challenges

Capacitance

Minimizing the traditional capacitance elements, such as under-lap capacitance (C_{xud}), channel capacitance, junction capacitances (both gated edge and area), and the inner and outer fringe capacitance, will become more challenging at the reduced dimensions of advanced technologies. Furthermore, in recent generations, gate and contact CD dimensions have been scaling slower than contacted gate pitch. This means that parasitic fringe capacitances (for example, contact-to-gate and epi-to-gate) are becoming significant issues.

Several classes of techniques will be important in controlling capacitances going forward. The first will be minimizing intrinsic capacitances (tailoring source-drain extension abruptness, source-drain to body junction engineering, etc.). The second will be minimizing parasitic capacitances (reducing the k of front end materials, reducing gate height, finding new etch-stop materials, etc.).

Resistance

Minimizing the traditional resistive elements such as the accumulation resistance (R_{acc}), and the spreading, silicide and contact resistances; will also become more challenging at the reduced dimensions of advanced technologies. Furthermore, resistive elements previously neglected (including interface and epi resistance) are becoming significant issues.

Several classes of techniques will be important in controlling resistances going forward. The first will be simultaneously improving R_{acc} and source drain junction depth (X_J) through source-drain engineering (laser spike anneal, plasma doping, co-implantation techniques, etc.). The second will be improving interface resistance through unpinning and controlling the Schottky barrier height (alloys, implant modifications to silicides, Schottky barrier source-drains, etc.).

Gate control

There are two major paths for maintaining short channel effects with continued gate dimension scaling. The first is by T_{ox} scaling (in particular through the use of HiK-MG). The second is by changing the transistor architecture to a thin-body architecture (such as a multiple gate FET, MUGFET, or an ultra-thin body, UTB, device).

HiK dielectrics deliver reduced gate leakage while enabling further T_{ox} scaling. The use of a metal gate (rather than polysilicon) eliminates poly depletion, resolves the V_T pinning issue seen with poly on HiK, and screens soft optical phonons for improved mobility [4]. However, the challenges with HiK-MG are also significant. HiK gate dielectrics contain a high level of traps and charge and thus pose significant challenges for reliability. A variety of scattering mechanisms result in reduced mobility. HiK-MG flows are complex, with the challenge of obtaining nearly band-edge workfunctions after thermal processing on both N and PMOS.

UTB and MUGFET devices deliver improved SCE performance, and enable further L_{eff} scaling [7]. However, these devices also have significant challenges. For MUGFET devices, the numerous free surfaces pose significant challenges in creating and retaining strain for performance enhancement. For UTB (and MUGFET) devices, the thin body (or narrow fin) produces significant resistance increase with associated degradation in drive current. As the body thickness is reduced to improve SCE, effects such as quantum mechanical V_T shifts and mobility degradation become significant. While undoped MUGFETS can provide mitigation for random dopant fluctuations (RDF); they also add new variation sources (most specifically, fin width, W_{si}, or body thickness, T_{si}, variation).

Mobility

Maintaining the scaling roadmap will require continual improvement in channel mobility. While advanced materials such as Ge or III-V materials offer potential long-term options, a shorter term approach for the 22nm or 15nm nodes may be reorient the surface or change the channel direction.

The best unstrained NMOS devices are fabricated on the (100) surface <110> direction, and the best unstrained PMOS devices are fabricated on the (110 surface) <110> direction [8]. Significant research in the last five years has focused on the challenge of trying to take advantage of the enhanced PMOS mobility on (110) <110> type material, without degrading the NMOS (for example, the HOT process, which integrates both orientations on the same wafer [8]).

Strain has had tremendous impact in advancing the transistor scaling roadmap [2,3,9]. A large number of process-induced strain techniques are employed in today's fabrication (e-SiGe, e-SiC, contact-etch-stop layer, stress-memorization technique, stressed gate metal, stressed contact metal, etc.). Future transistor architecture solutions (whether (100) or (110), planar or MUGFET) must possess significant strain enhancement on both N and PMOS to continue to drive the scaling roadmap forward.

Advanced channel materials (such as Ge and III-V materials) offer potential long term solutions for mobility enhancement when integrated with silicon substrates [10]. Unfortunately, the most interesting advanced channel materials are lattice-mismatched to silicon, with associated fabrication challenges. In addition, the low band-gap materials (InAs, InSb, Ge, etc.), display significant band-to-band tunneling (which may limit them to low voltage operation, or UTB implementation). Finally, use of these materials requires developing new gate dielectric fabrication techniques or moving away from MOS-style gate architectures.

Variation

The continued decrease in the ratio of feature sizes to fundamental dimensions (such as atomic dimensions and light wavelengths) means that management of variation will play an increasingly important role in future technology scaling [11].

The SRAM exercises the smallest area devices in the technology and thus SRAM V_{min} is the most critical parameter affected by random variation. Several classes of techniques will be important in controlling variation going forward. First, new patterning techniques will be needed to assure that patterning variation follows the historical scaling trends. In addition, improvements in doping methodology and gate control will be required to maintain or improve σV_T with further scaling.

3. Conclusions

While significant transistor challenges (capacitance, resistance, gate control, mobility, and variation) exist for technologies past 32nm, a variety of innovative solutions are being explored by researchers worldwide in the quest to drive Moore's Law forward.

References
[1] Dennard, R. et al., IEEE JSSC, Vol .9, No. 5, pp. 256–268, Oct 1974.
[2] Thompson, S. et al., IEDM Tech. Dig., pp. 61–64, Dec. 2002.
[3] Bai, P. et al., IEDM Tech. Dig., pp. 657-660, Dec. 2004.
[4] Mistry, K. et al., IEDM Tech. Dig., pp. 247-250, Dec. 2007.
[5] Natarajan S. et al., IEDM Tech. Dig., pp. 941–943, Dec. 2008.
[6] Auth, C. et al., 2008 Symp. on VLSI Tech, pp.128–129, June 2008.
[7] Liu T.J.K., Chang, L., "Into the Nano Era", Springer, Vol. 106, pg.191, 2009.
[8] Yang, M. et al., IEDM Tech. Dig., pp. 453-456, Dec. 2003.
[9] Rim, K., ICSICT 2008, pp. 105-108, Oct. 2008.
[10] Krishnamohan, T., Saraswat, K., ESSDRC 2008, pp. 38-46, Sept. 2008.
[11] Asenov, A., 2007 Symp. on VLSI Tech, pp.86-87, June 2008.