Moore's Law past 32nm: Future Challenges in Device Scaling

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

***Abstract*—This paper explores the challenges facing process generations past the 32nm technology node and speculates on what new solutions will be needed. The challenges facing planar and multiple-gate devices are compared and contrasted. Resistance and capacitance challenges are reviewed in relation to past history and on-going research. Key enhancers such as high-k metal-gate (HiK-MG), substrate and channel orientation, as well as NMOS and PMOS stress are discussed in relation to the challenges of the coming transistor generations.**

Keywords-CMOS; high-k; metal-gate; strain; orientation

I. INTRODUCTION

For the past 40 years, relentless focus on Moore's Law transistor scaling has provided ever-increasing transistor performance and density (Fig. 1). Interestingly enough, key technologists in each generation of this long history have also looked forward and predicted the "end of scaling" within one or two generations [1-4]. However, each time the technology reached the predicted barriers, scaling did not stop. Instead, imaginative new solutions were developed to further extend Moore's Law and the transistor scaling roadmap.

A decade ago, "transistor scaling" meant "classic" Dennard scaling [5], 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. Classic Dennard scaling ended at 130nm (130nm was the last CMOS generation where making the transistor smaller was sufficient to deliver performance improvement). In all subsequent generations (90nm, 65nm, 45nm, 32nm etc.) shrinking the transistor degraded the performance. However transistor scaling didn't end at the 130nm node. Instead, enhancers were added (strain in the 90nm and 65nm nodes [6,7], and strain + HiK-MG in the 45nm and 32nm nodes [8,9]) to continue to drive the transistor roadmap forward.

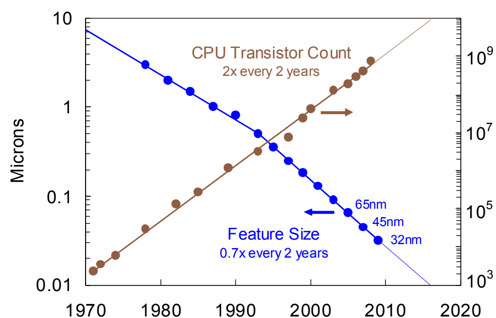

Fig. 1. Moore's Law: CPU transistor count has increased by 2X and feature size has decreased by 0.7X every two years.

II. NEXT GENERATION ARCHITECTURES

A. *Planar transistors*

As we look beyond 32nm planar transistors, there are a number of scaling challenges to be addressed (Fig. 2). Increased off-state current (I_{off}) from degraded drain-induced barrier lowering (DIBL) and subthreshold slope (SS) caused by poorer short channel effects (SCE) represents a significant limitation for effective gate lengths (L_{eff}) shorter than approximately 15nm. Decreasing T_{ox} to provide better channel control comes with a penalty of increased gate leakage current (I_{gate}) and increased channel doping. Increased channel doping decreases mobility (degrading performance due to impurity scattering) and increases random dopant fluctuations, RDF (degrading the minimum operating voltage, V_{min}). Decreasing gate pitch increases the parasitic capacitance contribution for both contact-to-gate and epi-to-gate thus increasing overall gate capacitance (C_{gs}). Decreasing source/drain opening size increases the source drain resistance (R_{sd}) thus decreasing drive current. Additionally, 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.

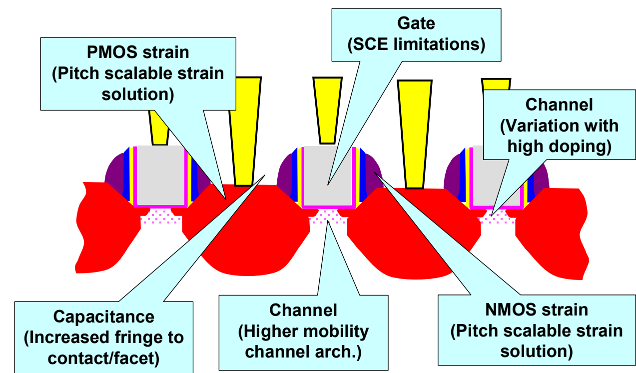

Fig. 2. Scaling challenges in planar devices.

B. *Multiple gate or multiple channel devices*

Multiple gate or multiple channel FETs (MUGFETS or MUCFETS) have been proposed to help resolve the SCE issues past 32nm (as an example, the ITRS 2007 roadmap predicts the use of multiple gate or multiple channel FETs starting in 22nm [10]). While MUGFETs provide significant resolution to the DIBL/SS SCE issues, they do have some unique challenges of their own (Fig. 3). The most significant of these is the challenge of continuing to maintain a high level of mobility

enhancement from stress in an architecture with numerous free surfaces generated by the fins. MuGFETs also require over-scaling the pitch to reduce capacitance and increase drive, (pitch over-scaling will be challenging on nodes still limited to 193nm immersion steppers). While MUGFETS may provide mitigation for RDF and thus provide V_{min} improvement; they also add new variation sources (fin width W_{si} variation in particular, is similar to L_{eff} variation in modifying the SCE properties of a device). Although C_{gs} and R_{sd} increases are a generic scaling problem for all architectures at tight pitches, the requirements for aggressive fin widths (to maintain SCE) and fin pitches (to meet drive current and capacitance targets) will enhance these problems in MUGFETS. Last, but not least, any 3'D architecture (such as a MUGFET) will require exquisite control of etch and patterning, and introduce topography risks for fill and polish.

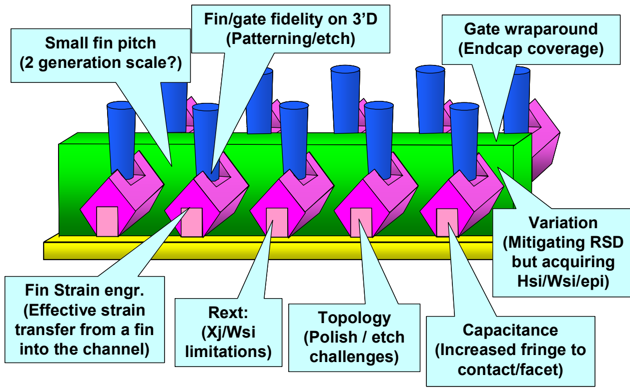

Fig. 3. Scaling challenges in MuGFET devices.

III. NEXT GENERATION CHALLENGES

A. Capacitance challenges

The traditional capacitance elements (Fig. 4), 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.

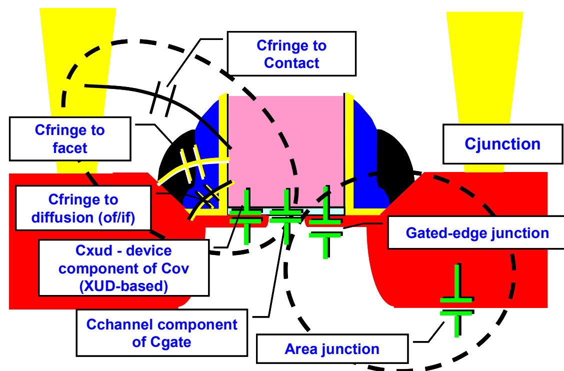

Fig. 4. Capacitance elements in planar architectures.

In addition to these elements, 3'D geometries (such as MUGFET devices) also introduce "dead space" parasitic capacitance associated with the region between the fins.

The key knob for parasitic capacitance improvement in the front end (for either planar or non-planar devices) is to reduce the k of the spacer as explored recently in the literature, for example in 2008 by Liow [11] and Ko [12].

B. Resistance challenges

The traditional resistive elements (Fig. 5), such as the accumulation (R_{acc}), spreading, silicide and contact resistances; will also become more challenging at the reduced dimensions of advanced technologies. Furthermore, resistance elements previously neglected (including interface and epi resistance) are becoming significant issues for planar devices. In addition, 3'D geometries (such as MUGFET devices) must compensate for additional resistance increases with decreasing fin width.

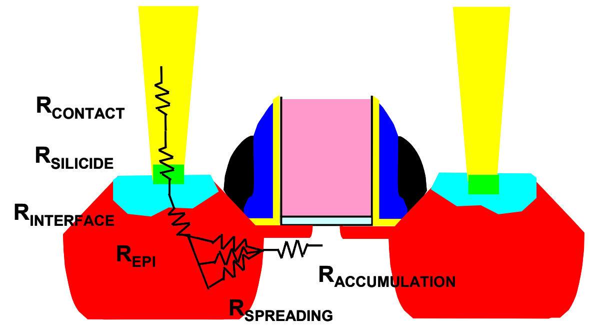

Fig. 5. Resistance elements in planar architectures.

Fig. 6 is a traditional scaling slide derived from the ITRS 2007 report which illustrates the increasing importance of R_{acc} in technology scaling [10].

A rich spectrum of techniques are being explored by a variety of groups for improving R_{acc} in a traditional planar architecture. Examples include laser spike anneal (LSA) as explored by Luo [13], Pouydebasque [14], and Yamamoto [15], non-melt LSA as reviewed by Ortolland [16] and advanced implantation techniques as discussed by Gelpey [17].

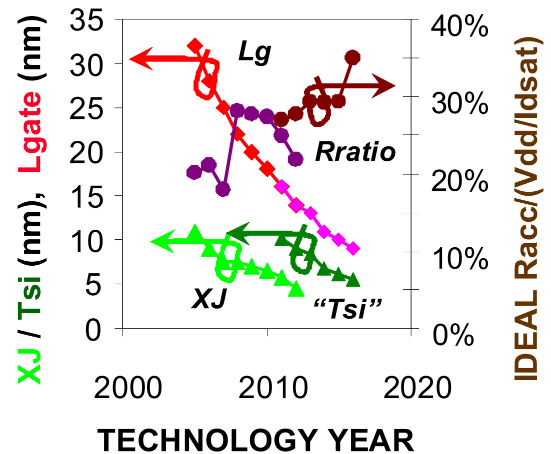

Fig. 6. ITRS scaling of X_J , L_{gate}, and R_{acc} [10].

Looking ahead to further resistance improvements, interface resistance improvement through modulation of the Schottky barrier height (SBH) offers significant opportunities. Theoretical SBHs span a wide range, with desired PMOS metals including noble metals (such as platinum and iridium) and desired NMOS materials including rare earths (such as erbium and ytterbium.) Unfortunately, the key practical challenge with improving SBH is that most materials on silicon pin the SBH at mid-gap. A variety of efforts are in progress to achieve theoretical SBHs through clever modifications to the process. One rich field of research is in alloy modifications to traditional silicides as discussed by Lee [18] and Ouchi [19]. Another area of critical study is implant modifications to traditional silicides, either on single metals or alloys as suggested by Zhang [20] and Larrieu [21].

C. High-k plus metal gate (HiK-MG)

HiK dielectrics deliver reduced gate leakage and enable 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 [22]. Mistry in 2007 confirmed these benefits on a manufacturable 45nm process demonstrating an 0.7X T_{ox} scale while simultaneously reducing gate leakage by 25X for NMOS and 1000X for PMOS with ring oscillator performance 23% better than 65 nm at the same leakage and 100mV lower Vcc [8]. Natarajan extended this work in 2008 showing 2nd generation HiK-MG results with 0.9A equivalent oxide thickness (EOT) and CV/I gate delay improved 22% compared to 45nm at the same leakage and Vcc [9].

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 (including soft optical phonon scattering [22]) result in reduced mobility. HiK-MG flows are complex; with significant challenges on thermal budget; and on the etch, polish and fill integration.

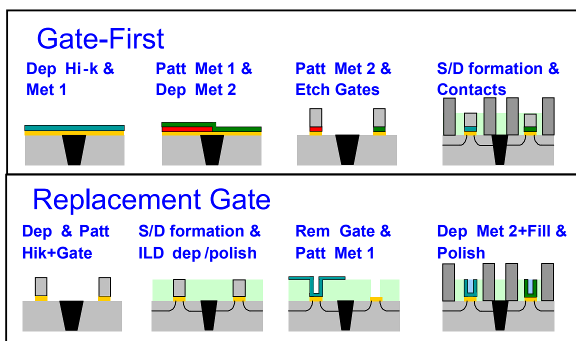

Fig. 7. HiK-MG flows, gate first vs. replacement gate.

The gate architecture choice remains a key point of discussion for HiK-MG. There are two primary competing architectures (Fig. 7), gate first (where both the dielectric and the metal are laid down before gate definition) and replacement gate (where a sacrificial gate is fabricated and later replaced by the metal gate material).

One benefit of the replacement gate flow is that it permits higher temperature anneals prior to the metal deposition (for better activation of implants). Another key benefit is that the replacement gate flow enables an elegant PMOS strain enhancement mechanism (Fig. 8) through first straining the PMOS with e-SiGe and then removing the gate [23].

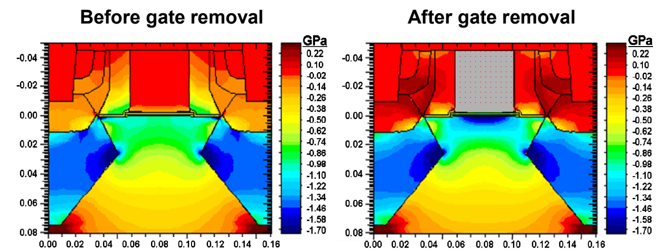

Fig. 8. Removal of poly gate increases channel stress by 50% [23].

Another key challenge in HiK-MG systems is obtaining nearly band-edge workfunctions after thermal processing on both N and PMOS. There are two approaches to this challenge; the one dielectric and two metals approach [8] or the two dielectrics and one metal [24].

D. Wafer and channel orientation

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 the channel orientation.

The classic orientation for silicon is a (100)-type surface; with two <110> channels perpendicular to each other and a <100> channel at 45 degrees. The competing orientation is a (110)-type surface with three possible channel directions; <110>, <111> and <100>. If both crystal orientations are possible simultaneously, the best unstrained NMOS is on the (100) surface <110> direction, and the best unstrained PMOS is on the (110 surface) <110> direction [25].

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. While it is theoretically possible to create both orientations on one wafer (by using a FinFET or vertical FET device on (100)-type material, and orienting the NMOS at 45 degrees, [26]) the 45 degree orientation poses significant challenges for lithography.

The alternative approach is to simply integrate both orientations on the same wafer. This process (termed HOT) was first reported by Yang in 2003 [27]. In this process, a wafer bonding technique is used to create SOI of the opposite type as the handle wafer for one of the transistor types. Epi from the handle wafer is then used to create a bulk transistor of the opposite type to the SOI. Sung in 2005 [28] introduced a direct silicon bond version of HOT that used bulk transistors for both N and P (rather than one SOI and one bulk). Yang in 2006 [29]; reversed the concept with a SuperHot process that integrates a pure SOI (100)-type and a pseudo SOI (110)-type device on the same wafer.

E. Strain

Strain has had tremendous impact on maintaining the transistor scaling roadmap. Because of the high gains provided by strain in today's processes, future transistor architecture solutions (whether (100) or (110), planar or MUGFET) will require significant strain enhancement on both N and PMOS. Reviewing the rich history of strain technology provides insight on what may be possible in subsequent generations.

Early work in CMOS strain followed examples from III-V systems, and introduced channel strain by epitaxial growth of lattice mismatched Si/SiGe systems. The seminal work in this area was done by Welser in 1992 [30] when he first quantified the strain enhancement in strained Si on relaxed SiGe. Hoyt in 2002 [31] expanded on Welser's work by exploring the strain enhancement with vertical effective field and doping . Also in 2002, Rim [32] further expanded on the Welser and Hoyt work by quantifying and controlling the lower V_T associated with short channel strained NMOS.

In 2000, Ito explored a different approach to strain engineering in CMOS [33]. He demonstrated a highly-manufacturable NMOS strain process by using the SiN contact etch stop layer (CESL), which introduced channel strain without the need to re-architect the channel. Pidin expanded on this concept in 2004 [34], by demonstrating simultaneous NMOS and PMOS CESL. Mayuzumi extended this to HiK-MG processes in 2007 [35], by illustrating that strain improvement was possible even with dual-cut CESL stressors.

On the PMOS side, Thompson in 2002 [6] presented a highly-manufacturable uniaxial PMOS strain solution (again with stressors outside the channel) using embedded SiGe (e-SiGe) material in the source/drain regions. Ghani expanded on this work in 2003 [36], and Chidambaram in 2004 [37]. This elegant technique caught on quickly and by 2005 many researchers were exploring this with representative examples including Lee with e-SiGE with SOI [38], Ohta illustrating the impact of e-SiGe profile engineering [39], and Zhang demonstrating e-SiGe on thin body SOI [40]. As an additional enhancement, Wang in 2007 [41] and Auth in 2008 [23] demonstrated that HiK-MG in a replacement gate flow enabled additional PMOS strain enhancement as a consequence of first straining the PMOS with e-SiGe and then removing the gate.

Stress memorization (where the gate is implanted, capped, then annealed to introduce channel stress) was an unexpected and highly successful stress enhancement technique. Ota in 2002 [42] first demonstrated the stress memorization technique (SMT) with Chen in 2004 [43] reporting the first large enhancements. More recently, Wei in 2007 [44] demonstrated that the process can be repeated multiple times for additional enhancement and Kubicek in 2008 [45] demonstrated SMT on HiK-MG.

There has been much recent interest in taking the PMOS enhancements with e-SiGe and applying the same techniques to NMOS. Ang in 2004 [46] first reported SiC enhancement of NMOS. Liu in 2007 [47], demonstrated SiC enhancement with implant plus SPE technique . Ren in 2008 [48], demonstrated SiC enhancement from in-situ epi plus P-SiC.

Last, but certainly not least, are the enhancements due to metal stress in the front end. Kang in 2006 [49] first reported enhancement from gate metal stress. Auth in 2008 [23] reported enhancement for gate metal stress and also reported enhancement with tensile contact stress.

IV. Conclusions

While significant transistor challenges (SCE, resistance, capacitance, mobility, etc.) exist for technologies past 32nm, many potential solutions are being explored to drive Moore's Law forward. As just a sampling of possibilities, 3'D architectures offer long-term solutions for SCE, SBH engineering and low-k dielectrics are being explored as solutions for resistance and capacitance challenges, and advanced strain techniques with alternative substrate/channel orientations are strong options for mobility enhancement.

References

[1] Broers, A. N.; IEDM Tech. Dig., pp. 2 - 6, Dec. 1980.
[2] Meindl, D.; IEDM Tech. Dig., pp. 8 – 13, Dec. 1983
[3] Wright, P. J. et al. IEEE TED, Vol. 37, pp. 1884 - 1892, Aug 1990.
[4] Heilmeier, G. H.; IEDM Tech. Dig., pp. 2 -5, Dec. 1984
[5] Dennard, R.H et al. IEEE JSSC, Vol .9, No. 5, pp. 256 – 268, Oct 1974
[6] Thompson, S. et al. IEDM Tech. Dig., pp. 61–64, Dec.2002
[7] Bai, P. et al. IEDM Tech. Dig., pp. 657-660, Dec.2004
[8] Mistry, K. et al. IEDM Tech. Dig., pp. 247-250, Dec. 2007
[9] Natarajan S. et al. IEDM Tech. Dig., pp. 941–943, Dec. 2008
[10] A. Allan, ITRS roadmap, 2007 ITRS Conf., Dec. 2007
[11] Liow, T.Y. et al. IEEE EDL, Vol. 29, Issue 1, Jan. 2008 pp.80 - 82
[12] Ko, C.H. et al. 2008 Symp. on VLSI Tech, 17-19 June 2008 pp.108–109.
[13] Luo, Z. et al. IEDM Tech. Dig., pp. 489–492, Dec. 2005
[14] Pouydebasque, et al. IEDM Tech. Dig., pp. 663–666, Dec. 2005
[15] Yamamoto, T. et al. IEDM Tech. Dig., pp. 143-146, Dec. 2007
[16] Ortolland, et al. Symp. on VLSI Tech; pp.186 – 187,17-19 June 2008
[17] Gelpey, J. et al. IWJT '08. pp.82 – 86, 15-16 May 2008
[18] Lee, R.T.P. et al. IEDM Tech. Dig., pp. 851-854, Dec. 2006
[19] Ohuchi, K. et al. IEDM Tech. Dig., pp. 1029–1031, Dec. 2007
[20] Zhang, Zhen et al. IEEE EDL,; Vol. 28, Issue 7, July 2007 pp.565 – 568
[21] Larrieu, G. et al. IEDM Tech. Dig., pp. 147–150, Dec. 2007
[22] Chau, R. et al. IEEE EDL, Vol. 25, No. 6, pp. 408–410, June 2004
[23] Auth, C. et al. 2008 Symp. on VLSI Tech; pp.128–129, 17-19 June 2008
[24] Cartier, E; Semicon West 2008, Session 8, June 25, 2008.
[25] T. Sato, Y. et al. Phys. Rev. B, vol. B4, pp. 1950-1960, 1971.
[26] Chang, L. et al. IEEE TED, Vol. 51, No.10, pp.1621–1627, Oct. 2004
[27] Yang, M. et al. IEDM Tech. Dig., pp. 453-456, Dec. 2003
[28] Sung, C.Y. et al. IEDM Tech. Dig., pp.225 – 228, Dec. 2005.
[29] Yang, M. et al. 2006 Symp. on VLSI Tech, pp.166–167, June 2006.
[30] Welser, J. et al. IEDM Tech. Dig., pp. 1000–1002, Dec. 1992
[31] Hoyt, J.L. et al. IEDM Tech. Dig., pp. 23–26, Dec. 2002
[32] Rim, K. et al. 2002 Symp. on VLSI Tech, pp. 98-99, June 2002
[33] Ito, S. et al. IEDM Tech. Dig., pp. 247–250, Dec. 2000
[34] Pidin, S. et al. IEDM Tech. Dig., pp. 213–216, Dec. 2004
[35] Mayuzumi, S. et al. IEDM Tech. Dig., pp. 293–296, Dec. 2007.
[36] Ghani, T. et al. IEDM Tech. Dig., pp. 978–980, Dec. 2003
[37] Chidambaram, et al. Symp. on VLSI Tech, pp. 48–49, June 2004
[38] Lee, W.-H. et al. IEDM Tech. Dig., pp. 61-64, Dec. 2005
[39] Ohta, H. et al. IEDM Tech. Dig., pp. 247-250, Dec. 2005
[40] Zhang, D. et al. Symp. on VLSI Tech, pp. 26–27, June 2005
[41] Wang, J. et al. Symp. on VLSI Tech, pp. 46–47, June 2007
[42] Ota, K. et al. IEDM Tech. Dig., pp. 27–30, Dec. 2002
[43] Chen, C.H et al. Symp. on VLSI Tech, pp. 56–57, June 2004
[44] Wei, A. et al. IEEE Symp. on VLSI Tech; pp. 216–217, June 2007
[45] Kubicek, S. et al. IEEE Symp. on VLSI Tech; pp. 130-131, June 2007
[46] Ang, K.W. et al. IEDM Tech. Dig., pp. 1069–1071, Dec. 2004
[47] Liu, Y. et al. Symp. on VLSI Tech, pp. 44–45, June 2007
[48] Ren, Z. et al. Symp. on VLSI Tech; pp. 172–173, June 2008
[49] Kang, C.Y. et al. Symp. on VLSI Tech, pp. 885-888, Dec. 2006