BetaSights℠ Newsletter
Commercial Free • Member Supported • Productive Info
www.betasights.net

Volume 00 Issue 01, 2009 January 1

Contents:

BetaSights Newsletter

Founder and Editor
Ed Korczynski
edk@betasights.net

Litho & DFM Editor
M. David Levenson
mdl@betasights.net

Managing Editor
Elizabeth Schumann
elizabeths@betasights.net

BetaSights *Newsletter* is published 44 times a year for the supporting Members. People provide information about technologies to BetaSights, so that edited articles can appear in the Newsletter. There is never a charge for information to be published by BetaSights. To submit info, send an email to info@betasights.net.

Welcome to BetaSights

Peer-review organizations such as the IEEE and MRS cover advanced R&D, while consultancies cover mainstream manufacturing business and market dynamics. However, before an IC, FPD, MEMS, or PV technology can move from alpha R&D into commercial manufacturing, it must be evaluated at a fab beta site to prove fit for an application. It is these critical materials, equipment sets, and services that BetaSights tracks.

Ed Korczynski worked with 50μm thin silicon for TSV in 1999.

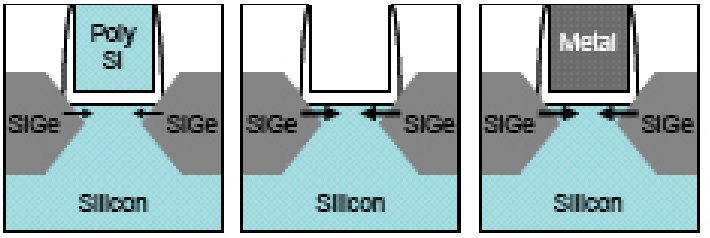
The www.betasights.net home page provides a regularly updated Sights table (still under construction at the time of this publication) of beta information. Basic information is free for anyone to access without registration, while additional in-depth information is available for Members-only access.

Weekly BetaSights *Newsletters*—such as this one—are published online at www.betasights.net for viewing and downloading by members. In addition to full *Sights* table access (as described above), Members have the ability to comment on BetaBlog℠ postings, as well as access to Members-only forums and an online calendar. Membership is set up online from the BetaSights home page, and currently costs only $79/year per person. –E.K.

IEDM2008 Trend-setters united

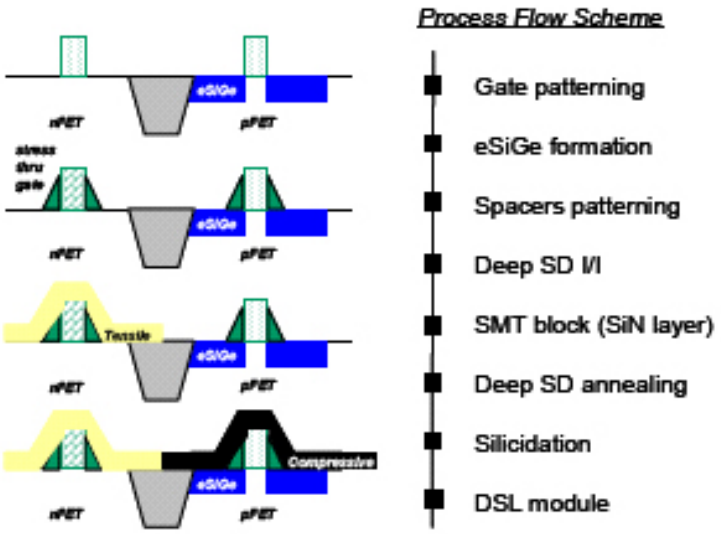
The 2008 International Electron Devices Meeting (IEDM) of the IEEE occurred mid-December in San Francisco, and ~1,400 of the world's top technologists gathered to showcase their latest, greatest processes and structures. While Intel and IBM continue to compete in 2D shrinks of planar CMOS circuits, TSMC continues to be the fastest follower in fabdom.

All three companies are united in asserting that the trend of two-years- per-node 2Dshrinks will continue for now...with the 32nm node expected 2H09-1H10. Intel's C. H. Jan, Logic Technology Development, Manufacturing Technology Development, claimed that 1/2 of all transistors sold today by Intel use High-K/Metal-Gate (HKMG) technology.



Intel reported on its latest 32nm results [Session 27.9] using a replacement gate HKMG flow (*see Figure, above*) and immersion lithography to get to 112.5nm contacted gate pitch. Intel asserts that the transistors feature 4th-generation strain technology, along with, "9Å EOT gate dielectric," which may be the thickness just for the hafnia-based High-K portion of the stack. At 1.0 V_{dd} and 100nA/μm I_{OFF}, NMOS and PMOS saturated drive currents were reported as 1.55mA/μm and 1.21mA/μm, respectively.

IBM's alliance of ST, Freescale, Chartered, Infineon, Samsung, and Toshiba presented on 32nm general purpose bulk CMOS [Session 27.3] with 0.126μm poly contacted pitch and 28nm Lpoly, without needing silicon-on-insulator (SOI) wafers (*see Figure, below*). Using SOI reduces fab costs and variability, but adds wafer cost. Separately [Session 27.8], Toshiba showed single-exposure litho to limit 32nm bulk costs.



Computer peripherals, graphics systems, networking servers, and digital consumer devices all require high data rate transfers at low voltages to save power. Since Chartered and Samsung are both pushing foundry business models, and since ST, Freescale, and Infineon all fab SoCs, a low-cost 32nm process is essential for their fab needs.

With "1st-gen" HKMG chips in production at 45nm, the industry is now working on "2nd-gen" improvements to meet 32nm and 22nm node targets. SELETE provided [Session 2.5] a thorough overview of different oxides that could be used as bases-below and caps-over first generation HK materials, including base/capping Y_2O_3, capping La_2O_3, and capping MgO_2. La_2O_3 cannot be removed by 0.25% DHF, while MgO_2 and Y_2O_3 layers can be easily etched by this DHF.

Deposition of these new materials is done using atomic-layer deposition (ALD) chambers and the following precursors and conditions:

- $Mg(EtCp)_2$ precursor ~35°C, Ar carrier gas, 30 sccm; and
- $La(i\text{-}PrCp)_3$ and $Y(TMOD)_3$ precursor 150°C, Ar carrier gas.

V_T is not effectively controlled with capping MgO_2 layers. For capping-La_2O_3 layers, just one monolayer provided ultra-thin EOT (0.72nm) with ΔV_T <130mV, and very high drain current (>1100μA/μm) at a low I_{OFF} (100 nA/μm). From EELS/TEM analysis, La atoms are ~6% in the HfO gate stack capped by La_2O_3.

SEMATECH's Paul D. Kirsch [Session 2.6] presented on "Device and Reliability Improvement of HfSiON+LaO/Metal Gate stacks for 22nm node applications." Evidence now shows that La diffusion in HKMG stacks is the critical parameter behind the observed as much as 500mV of ΔV_T occurring with interfacial dipoles when using dielectric caps. The source of these dipoles is the LaO cap material diffusing through the hafnia; metrology data show that both increasing the thickness and nitriding the hafnia layer reduces the formation of these dipoles.

With the understanding that La diffuses easily, and with the goal of La at the bottom of the hafnia (but not reaching the Si surface) to tune the dipoles, nitridation of the "native oxide" is used as a diffusion barrier. Using SiON below the hafnia retards the La diffusion past the interface, keeping the La away from the channel.

For low V_T and low EOT CMOS integration, there are three basic approaches:

1. Dual metals,
2. Dielectric caps, and
3. Dual channels (such as SiGe).

For 22nm node manufacturing there are no simple solutions, only complex trade-off decisions, and so the choice of which of the three approaches to take is based on performance/cost estimates for target applications. For LSTP applications of digital CMOS, the current flavor is an AlO cap for NMOS (~1.2 EOT) and LaO for PMOS (~1.5 EOT) with a single metal.

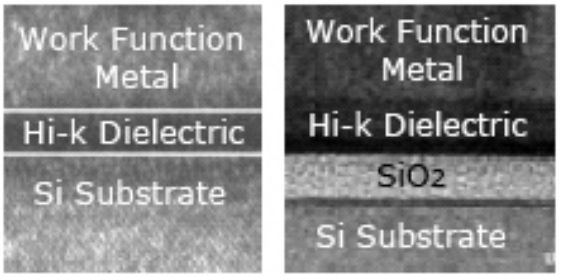
"The lanthanum oxide is very flexible. We're able to design EOTs over quite a wide range from 1.4 for LSTP down to 0.8 for HP applications," explained Kirsch. HP pFET technology uses gate-first epi-SiGe for both low EOT and VT. HP nFET technology uses a LaO cap and the same metal-gate (so only one metal needs to be etched), and in 30nm gate length with a strained etch-stop and spacer provides 1200μA/μm.

In the "Future of Fabs" evening session, Craig Sander, corporate vice president of process R&D, The Foundry Company, provided the perspective, "Is the technology node progression slowing down? Not yet, but it soon will." Certainly, power/performance advances are more difficult with each node, and development costs continue to grow exponentially (up to 1.4x/node). Beyond 22nm for logic, increases in processed wafer costs could fully offset any density advantages.

"Innovation isn't beginning to slow down," explained Sander. "Innovation beyond device scaling is just getting started. We've been talking about finFETs for years, but it will first really start at the 22nm node." We've talked about 3D for decades, but the mainstream is only now starting to consider doing something non-planar. *–E.K.*

IEDM2008 Intel 45nmHKMG SoC

Having invested in the replacement metal gate (RMG, a.k.a. "gate-last") process flow for high-performance (HP) at the 45nm node, Intel has now developed HKMG at 45nm for low power SoC [Session 27.4] applications. Though the rest of the world has chosen "gate-first" HKMG transistor technology, Intel asserts that gate-last is not just a specialty process but suitable for mainstream use.



SOC Technolgy building blocks include the following: SRAM/RF, inductors, analog/HV/IO, decap, varactors, diodes, precision linear capacitors, in addition to HP and low power (LP) logic transistors. LP transistors must leak at least 100x less than HP, and use 40nm gate length while HP uses 35nm.

Standard 45nm node HKMG transistors can support up to 1.8V operation, but SoCs require input/output transistors that can run at 3.3/5.0V and so Intel has an integrated flow that adds an oxide layer below the HK (*see Figure, above*) for the I/O portions of the circuit. To survive 2.0-2.5V, the I/O uses L_g 160nm min. ~400 Ghz RF function has been shown using 160nm minimum pitch.

RMG flow begins as follows:

- Isolation,
- Oxide I/O gate dep. and patterning,
- HK logic gate dep.,
- Poly Si dep. and patterning, etc.

Using embedded SiGe contacts, raised S/D for PMOS, and "3rd-generation" strain engineering for both PMOS and NMOS, the SOC process has shown the same low defect density as Intel's CPU process and has been qualified in two fabs. *–E.K.*

IEDM2008 Remember new ways

Stefan Lai, the "Flashy" ex-Intel guy now touting Phase-Change Memory (PCM) technology at Being Advanced Memory Corporation, provided a keynote overview of Flash and other non-volatile memory (NVM) chip possibilites. All leading memory companies presented on NVM or DRAM or both over the three days of parallel sessions.

The world of materials for memory has become very interesting. From DRAM to SRAM to Flash, the world has moved from simple oxides and nitrides to more exotic binary and trinary oxides. Since a binary bit may be stored by any distinguishable change of state in matter, memory cells may theoretically be made out of almost anything. "Technologies may have unique applications and markets, but the lowest cost per bit is what results in the largest market," explained Lai. Consequently, magnetic RAM (MRAM) is among the technologies dismissed by Lai as inherently niche.

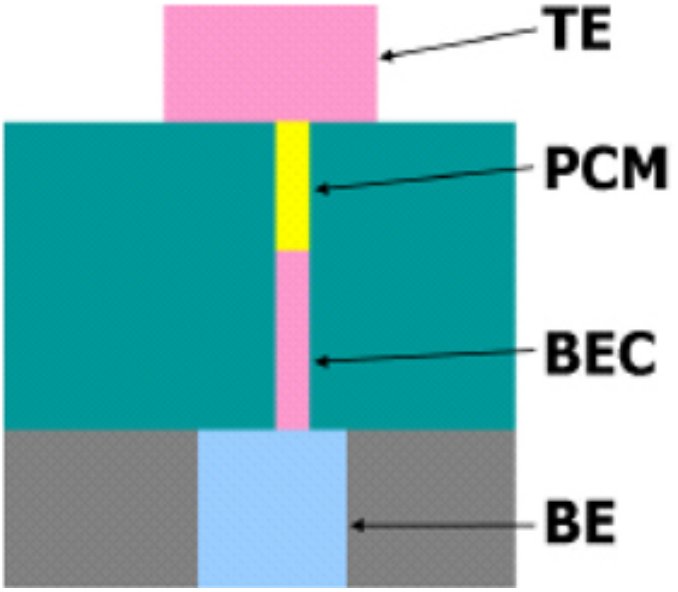
Key NAND cost reduction drivers have been lithography, the use of self-aligned techniques, and multi-level cells. Flash price—not cost—reduction has been due to relentless innovation and competition, and has driven flash memory price down to $1/Gbyte, which may be a tipping point for even more widespread use (beyond USB "thumb" drives).

> "It was one of the mistakes of my career to say that NAND doesn't work."
> -Stephen Lai

NOR and NAND scaling will continue for the next few years through continuous innovation and improvement. Despite concerns with severe-sub-wavelength lithography, 22nm Flash should be fairly do-able with regular layouts.

However, one fundamental limit is the number of electrons needed to store a distinct state in Flash, even using hemispherical and floating-trap storage ideas. "When I began working, I had thousands of electrons to start with, but now if you're losing one electron out of twenty it won't take you very long to lose information," reminded Lai. Distinct states could be lost, and multi-level capability would fail.

To compete with Flash, any new memory will have to show multi-level programming to acheive attractive cost/bit. Multi-layer structures may also be needed, where a new memory technology is combined with DRAM, SRAM, or Flash. It is for all these fundamental reasons that Lai advocates for phase-change memory (PCM) using some variation on chalcogenide.



Samsung seems to have pushed PCM technology the farthest [Session 9.2], showing a CVD process for the formation of the phase-change material into a 8.5nm wide "dash" confined cell (*see Figure, above*). The confined structure limits the PCM reset current to just ~160μA, and excellent reliability is claimed. The company has also demonstrated a 512Mb chip with direct write E10 cycles and switch time of 50ns.

"NAND broke the memory business paradigm by saying that you can have bad bits, and the controller can detect and correct any errrors. It was one of the mistakes of my career to say that NAND doesn't work," confessed Lai. With an overall system-level perspective, Lai sees that the write bandwith limitation of PCM can be handled with a DRAM buffer ~10% of the PCM size.

IBM showed [Session 27.1] what may be the limits of planar CMOS with a 22nm 0.1μm2 6T-SRAM, done with double-exposure, double-etch, co-implant and thin silicide, with Cu contact plugs. Beyond this, even the models start to break down.

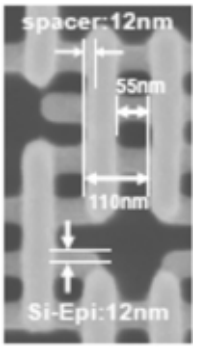
In the IEDM press lunch, IBM's Xiaomeng Chen questioned, "There are ways to get to smaller dimensions by etching or growth, but what will the final atomic structure be?" Published results from tests of carbon nano-tubes (CNT) as the functional elements of memory cells to date have been based on CNTs that were simply spun onto the substrate, but what are the charges flowing through the substrate verses flowing through the CNT? *–E.K.*

IEDM2008 finFET fundamentals

Mismatch gets worse when technology scales for planar bulk transistors. As the channel dimension is reduced the inherent variabily in physical parameters creates mismatch. Compared to planar, better drive strength with finFETs means improved read currents. The nature of mismatch in finFETs [Session 10.3] is radically different compared to bulk, and gain mismatch can dominate SRAM yield effects.

SOI finFET basic process flow:

- Fin patterning (i193nm),
- Gate stack deposition (HfSiO, TiN),
- Gate patterning, and
- Extensions (Spacers, Selective Epi).



It appears that fin width doesn't affect $V_{T,lin}$ for relevent gate lengths and widths. With fin structural optimation, gain mismatch in finFETs can be kept below the 10% target for SRAM arrays. Static Noise Margin (SNM) tests show all chips readable below 1V, with one optimized cell readable to 0.4V. Excellent SCE control and undoped channels result in σ-ΔV_T of 20 mV.

Toshiba/IBM/Freescale/AMD presented [Session 10.2] on SRAM arrays made using 30nm gate length HKMG finFETs in 0.187μm^2 cells. They also created and showed images (*see Figure*) of what was claimed as the world's smallest non-planar SRAM cell (0.128μm^2). In addition, the unit process steps and integrated flow all show scalability.

28nm high and 56nm channel width was assumed for the finFET, while the planar FET was modeled as 70nm pitch and Lg=24nm. ΔV_T of transistors in 0.187μm^2 cells was measured with and without channel doping. Using 22nm node design rules, an un-doped finFET SRAM cell was simulated to have significant advantage in read/write margin over a planar-FET SRAM cell, which would have higher ΔV_T mainly caused by heavy doping into the channel region. *–E.K.*

IEDM2008 Constant variability

In an invited paper [Session 17.4], Prof Asenov, University of Glasgow (http://www.elec.gla.ac.uk/groups/dev_mod/presentations.php has all of the slides) described recent advances in predictive physical simulation of statistical variability using drift diffusion (DD), Monte Carlo (MC) and quantum transport (QT) simulation techniques.

Statistical variability comes from the discreteness of charge and the granularity of matter in transistors with features already of molecular dimensions. Two transistors next to each other on the chip with exactly the same geometries and strain distributions may have characteristics from each end of a wide statistical distribution. 30mV of ΔV_T due to random traps have been seen.

Statistical variability restricts supply voltage scaling, which adds to power dissipation problems in SRAM designs, and to timing problems and hard faults in logic circuits. Changing from doped poly-silicon to HKMG helped reduce variability for one or two nodes, but HKMG structures depend upon grain sizes and orientations. "If you develop your 22nm node technology, and it turns out that you have too much variability and designs don't work," hinted Asenov, "it will be very bad."

Non-equilibrium Green's Functions are needed to try to model quantum effects (confinement, tunneling, etc.) in nanowires and dots. "It's much more difficult compared to drift modeling or Monte Carlo," according to Asenov. "Nanowires may be very interesting from a research point of view, but we have to really think about how much variability they will introduce."

Some companies may be happy with 10% yield on high-performance chips. Statistical variability implies new approaches to design.

It is likely that the statistical variability in bulk MOSFETs will increase rapidly in the next nodes, regardless of attempts to fine-tune the process. FD-SOI and dual-gate devices have significant advantages in terms of variability compared to bulk MOSFETS. *–E.K.*

Calendar

SEMI Industry Strategy Symposium, Half Moon Bay, CA, Jan 12–14

SEMI Strategic Materials Conference, Half Moon Bay, CA, Jan 14–16

Levitronix CMP & Ultrapure Fluid Users Conferences, Santa Clara Convertion Center, Feb. 10–11

SPIE Advanced Lithography Conference, San Jose Convention Center, Feb. 23–27