



White Paper: MST® Benefits DRAM Sense Amplifiers

Robert Mears & Jeff Lewis, Atomera Incorporated

Outline

In this White Paper we outline some recent learning about Atomera's MST® technology and how it can be used to improve the mismatch inherent in DRAM sense amplifiers. This is becoming a critical issue for the latest DRAMs because bitcell capacitors are hitting their scaling limits and the total stored charge is now declining. Reducing mismatch enables numerous cost- and power-reduction benefits that will be described.

Table of Contents

- Introduction
- Description of DRAM Sense Amplifier Mismatch
- Causes of Sense Amp Mismatch
- How MST Significantly Reduces Sense Amp Mismatch
- Doping profile and RDF: MOS 101
- MST and the ideal profile for minimizing RDF
- Benefits of Reduced Mismatch
- Summary
- References

Introduction

We call our technology MST, which stands for Mears Silicon Technology. MST is “quantum-engineered” to enhance transistor performance, based on over 15 years of R&D. In our previous White Papers (“MST® for PMIC and RF SOI Switches” and “Smoothing the Surface: MST® Turbo-Charges GAA Devices”) we focused on the proven ability of MST to engineer doping profiles and thus enhance power switch devices and our recent discovery that MST improves surface roughness scattering mobility.

In this paper we focus on using MST in a DRAM sense amplifier to significantly reduce sense amp mismatch. This issue has become critical as continued DRAM scaling is reducing the storage capacity of DRAM bitcell capacitors and increasing the economic pressure to shrink the sense amp transistors.

Description of Sense Amplifier Mismatch

DRAM (Dynamic RAM) chips operate by storing a charge on the capacitor of each bitcell. A high voltage often (but not always) means a ‘1’, while no charge (ground) means a ‘0’. These memories are dynamic because the charge on the capacitor leaks away over time, meaning the bitcell must be periodically refreshed. For example, to guarantee the integrity of all the bit cells

MST® Benefits for DRAM Sense Amplifiers

on a DDR5 products at 85 deg C, 8,192 refresh commands need to be issued every 32 ms. This translates to one refresh command every 3.9 us, or 256,000 refreshes per second.

Reading the value in the bitcell is the job of the DRAM sense amplifier (sense amp). A DRAM read operation starts by accessing a row of bitcells where each bitcell in the row shares its charge with a bitline. This causes the bitline voltage for each bitline to rise or fall, depending on the charge stored in its respective bitcell. A sense amp is coupled to each bit line and compares the voltage on the bitline to a reference bit line precharged to a reference voltage (V_{ref}). It amplifies the sensed signal, driving each bitline to a full '1' or '0,' depending on whether the bitline was higher or lower than its reference bitline.

This is the ideal case. In reality, every sense amp has a margin of error (called sense margin) that determines the minimum allowable difference between the bitline voltage and V_{ref} . The best possible sense margin would be 0V, but this is impossible to achieve. If the difference between the bitline voltage and V_{ref} is less than the sense margin, the sense amp cannot reliably detect the data value.

Sense amps have a sense margin because the transistors within them do not perfectly match, especially in threshold voltage (V_t), which tends to vary randomly around a target value. This stochastic variability is measured by the parameter σV_t ("sigma- V_t "), also called "local V_t mismatch". σV_t is typically plotted against the reciprocal of the square-root of the channel area of the device on a Pelgrom Plot, as shown below.

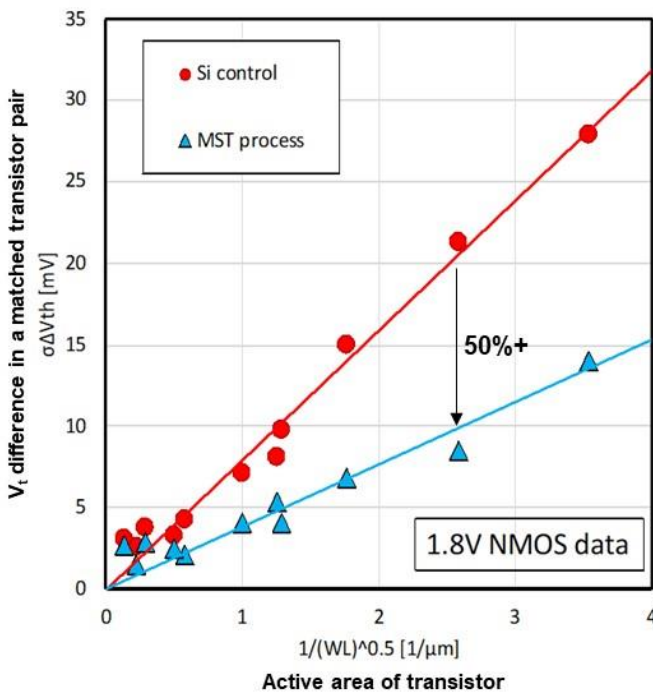


Figure 1: This is a Pelgrom Plot showing that MST has reduced the V_t mismatch in this silicon example by 50%. Also note the correlation of smaller channel area (moving to the right on the X-axis) leading to greater V_t mismatch between devices.

Causes of DRAM Sense Amp Mismatch

- **Random dopant fluctuation (RDF):** This is the biggest contributor, typically responsible for 50% or more of total σV_t . RDF exists because dopant deposition and diffusion are inherently stochastic processes: there is variation in how many dopant atoms each transistor receives in the channel region, and this is further exacerbated by dopants diffusing into and out of the channel during processing. Note that this problem gets worse with scaling because of the smaller total device area and number of dopant atoms per device, meaning one more or less dopant atom will have a bigger impact on the V_t value. As will be explained in the next section, MST can tremendously lower the σV_t from RDF.

- **Line Edge Roughness (LER):** As its name implies, LER is the deviation (when looking top-down) of the feature edge of a device from a smooth, ideal shape. Random (as opposed to systematic) LER is typically caused by the stochastic behavior of the manufacturing process.
- **Gate Oxide Thickness (GOT) variation:** Also, as its name implies, gate oxides are not perfectly uniform across the device; thinner/thicker spots will cause V_t to decrease/increase, respectively.
- **Metal Gate Granularity (MGG):** Only a factor for devices with High-K Metal Gates (HKMG), which are only just now being used in DRAM sense amps.

How MST Significantly Reduces Sense Amp Mismatch

It is well known in semiconductor physics literature that the vertical and lateral doping profile of a MOSFET plays a significant role in determining the amount of random dopant fluctuation, RDF [1].

Thus, the design of the periphery sense amp transistors (M1-M4) needs to take RDF into account. Unfortunately, it is difficult to optimize the doping profile and hence minimize RDF because of the various high temperature steps in a typical DRAM manufacturing process. The problem only becomes more acute as the periphery devices are shrunk in successive DRAM generations. As we saw in Figure 1, the magnitude of the mismatch depends inversely on the channel area of the transistor. Smaller transistors have greater mismatch. Also, to maintain control (I_{on}/I_{off}), the channel doping is increased as devices shrink, and this also leads to higher RDF and hence greater mismatch.

Insertion of MST during the front end of the sense amp transistor formation and optimization of the implants can create unique doping profiles which can minimize RDF and hence significantly improve mismatch.

To understand why, we first look at the impact of the doping profile on RDF and then describe how MST can be used to facilitate a near optimal doping profile.

Doping profile and RDF: MOS 101

A simplified profile of a planar NMOS transistor is shown in Figure 3. When a positive potential is applied to the gate, the positive charges on the gate repel the holes in the p-type body, leaving the negatively charged acceptor atoms (in this case boron). The electric field from the gate penetrates the body and reduces each time it impinges on the negatively charged boron atoms until a depth where the field falls to zero. This depth is known as the depletion width. The potential barrier, ΔV , for a hole to travel from below the depletion width to the semiconductor surface is found by integrating the position dependent field, since $E = -dV/dx$ (in

one dimension, where x is the vertical coordinate from gate toward depletion width) or more formally $\mathbf{E} = -\text{grad } V$. The magnitude of this potential barrier is represented by band-bending on the typical representation of the band diagram for the p-body.

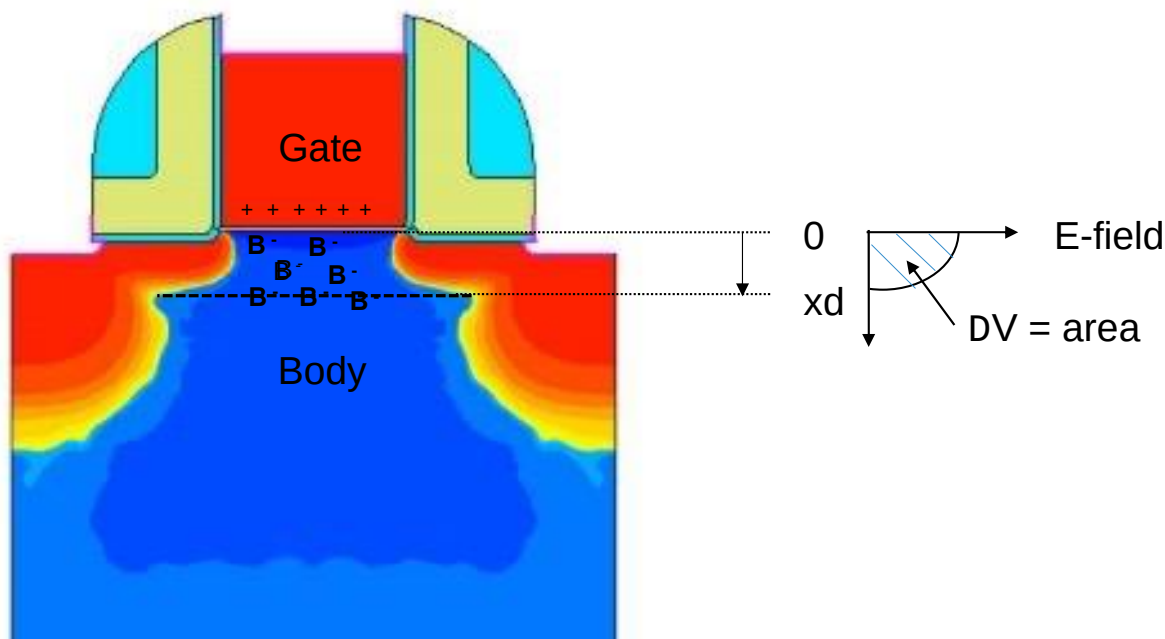


Figure 3. Simple schematic of NMOS device. Positive charge on gate (positive potential) repels positive holes in p-type body, leaving negatively charged acceptors (boron). In the silicon body, the E-field extends to a depth x_d (depletion width) where it falls to zero, once the encompassed negatively charged boron atoms balance the positive charge on the gate.

At the depletion width the p-body has a Fermi level close to the valence band, according to the p-doping at that depth. Because the Fermi level is constant in equilibrium, the band-bending means that the p-body is less “p-type” towards the surface because the local Fermi level moves away from the valence band. As the gate voltage is further increased, the amount of band bending increases until the surface of the p-body becomes n-type (the Fermi level is closer to the conduction band). We describe that this region near the gate as being in inversion. Current (electrons) can now flow more freely between the n-type source and drain. Although precise definitions of threshold vary, we generally describe this voltage as the threshold voltage, V_t . Further increase of the gate voltage (more charge on the gate) is now balanced by the electrons flowing in the inversion layer and the depletion width no longer needs to increase.

Without doing the full analysis, the key physics behind RDF can be explained by considering two identical NMOS transistors with the same average dopant distribution and thus the same nominal V_t . Let’s imagine that they are biased at V_t and the depletion width is x_d .

But because of statistical variations in the manufacturing process, let's imagine that transistor A is missing a boron atom at depth x but otherwise has an identical distribution to transistor B.

To first order, the Electric field doesn't care where the negative boron atoms are, it just checks them off until the number balances the number of positive charges on the gate, by which point it has fallen to zero. The field in transistor A needs to find one more charge, to make up for the missing boron, so the depletion width must extend deeper by a fraction. Usually at that depth there are plenty of boron atoms around so that fractional increase is negligible compared to x_d . Note that the E-field in transistor A relative to transistor B increases by an amount δE at depth x where the boron is missing, and then finally hits zero once it finds the extra charge at the depletion width.

The potential, on the other hand, which determines the band bending, is the integral of the E-field from the depletion width towards the gate. So, in transistor A, the additional band-bending, δV is $\delta E \cdot (x - x_d)$.

This equation is the key to understanding RDF. The additional band-bending in transistor A means that it is now in stronger inversion compared to B. In other words, the effective threshold voltage of transistor A has been lowered by an amount $\delta E \cdot (x - x_d)$. The magnitude of the threshold difference depends on the position of the missing charge. The threshold difference is maximum if the charge is missing at the surface ($x=0$), falling to zero if the missing charge is at depth $x=x_d$. So, RDF is highly sensitive to the position (depth) of the charge variation.

MST and the ideal profile for minimizing RDF

Why not skew the doping profile so that it is low at the surface and higher at x_d ? This is called a retrograde profile. It is well-known in the literature to reduce mismatch from RDF [1]. In fact, if the doping can all be placed at (or practically just above) the depletion width (delta-doping) then, as we saw above, RDF falls to zero!

Unfortunately, there are practical challenges to creating such profiles as devices are scaled. The first is that the regular diffusion processes from the thermal steps in the manufacturing process, and particularly from oxidation steps, tend to flatten the as-implanted well-doping profile. While retrograde profiles are possible in older, larger geometries, the depletion width is only a couple of tens of nanometers in typical sense amp transistors, and over such small distances, if anything, the doping profile tends to increase towards the gate, making RDF worse. The second problem is that as the gate length is reduced, additional doping is required to reduce off-state leakage and maintain good control of the device. This doping is normally achieved by insertion of halo doping after gate formation. This is a shallow implant at the surface of the device – just where you do not want additional doping from an RDF perspective!

The problem is illustrated by showing the vertical doping profile of a transistor of similar dimensions to the current generation of sense-amp transistors. The control silicon device shown as the blue plot in Figure 4. has an almost flat doping profile between 0nm at the surface of the device to 20-40nm, which would be a typical depletion width. In fact, the profile increases slightly towards the surface – the exact opposite of what is needed to reduce RDF.

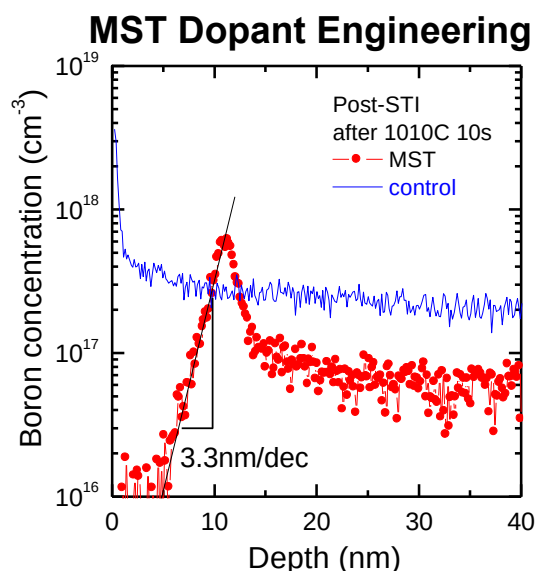


Figure 4. MST enables dopant engineered profiles that are robust to the thermal steps of modern semiconductor manufacturing processes. MST enables a “quasi delta-doping profile” which can be positioned very close to the depletion width, minimizing RDF.

The MST oxygen layers can be used to create a doping profile that is more than an order of magnitude lower at the surface and peaks at a depth determined by the positioning of the MST layers. This depth can be tuned in the MST deposition process according to the sense-amp transistor design, so that it lies just above the depletion width of the resultant device. Note that the plot is logarithmic, so that the MST doping profile is a very good approximation to the ideal delta-doping profile which reduces RDF to zero.

But what about halo-doping? If the process still requires a halo step to control leakage, then the MST RDF benefit above is thrown away. Fortunately, the MST profile can also be engineered to reduce the need for halo doping. It is well known, for example, that “super-steep” retrograde profiles significantly improve short-channel effects leading to less requirement for halo doping. In fact, we have shown that MST transistors can be engineered down to 22nm gate length (much shorter than current sense-amp transistors) without the need for a halo implant [2]. So,

with proper design, the sense amp transistor can be designed using MST to have low off and minimal RDF.

In one of those rare conjunctions of nature, the benefit does not stop there. By removing the charges close to the gate dielectric (and, as we discovered recently, also reducing the surface roughness scattering) MST increases the mobility of electrons and holes, giving higher currents just above threshold and also in saturation. For electrons (NMOS), this benefit was characterized and published with co-authors from UC Berkeley and SK hynix [3] as is shown in Figure 5.

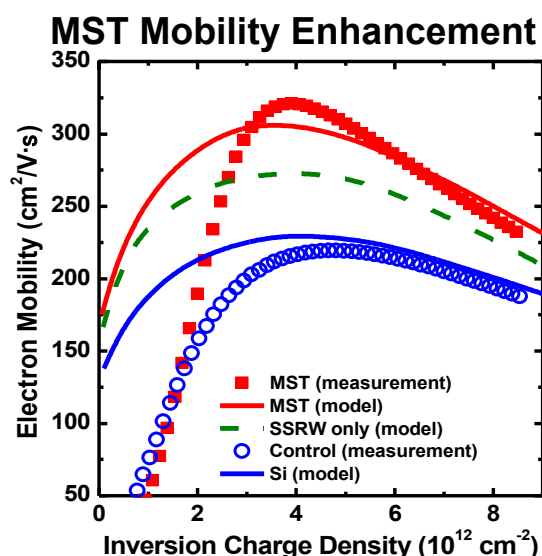


Figure 5. By moving the well doping away from the dielectric interface and by reducing surface roughness scattering, MST enables peak and high E-field mobility enhancement [3].

The MST mobility benefit has been further characterized in joint publications with Prof. Suman Datta and team [4] (for MST plus HKMG dielectrics) and over a wide temperature range, down to cryogenic temperatures, with Prof. H Y Wong [5], elucidating the improved surface roughness scattering mobility for both NMOS and PMOS devices.

Benefits of Reducing Mismatch

Sense margin has a profound effect on many facets of DRAM design. In fact, improving the sense margin is like “money in the bank” that can be one of the best ways to reduce product costs, lower power, and/or improve the performance of DRAM devices. Keep in mind that virtually every DRAM parameter has a stochastic distribution, so all design decisions must be made at the worst-case levels (or nearly so) to prevent data corruption.

A better (smaller) sense margin could improve any of these parameters:

- **Longer DRAM refresh interval:** DRAM bitcells must be refreshed when the charge on the (worst) capacitor leaks sufficiently that a '1' cannot reliably be read. Refresh consumes significant power – some studies show it consumes 30% of the total DRAM power budget, and 10-13% of total server power. A better sense margin enables a longer interval between refreshes, reducing power proportionally, possibly by 50% or more.
- **Cheaper, lower-complexity DRAM capacitors:** This is the flip side of the previous item. For many years, DRAM makers were able to shrink the die area of the capacitor and maintain the same capacitance value by growing the capacitor vertically and using exotic materials – resulting in DRAM capacitors with aspect ratios of 50:1 or more. But continued scaling has hit limits – the capacitance value is now dropping and therefore refresh intervals are getting shorter. The solution is to introduce even more exotic/expensive tricks to maintain cell capacitance, stop scaling the bitcell, or, ideally, improve the sense margin so that lower-capacitance bitcells can be used with similar or even longer refresh intervals.
- **More bitcells per bitline:** This is a little more complicated. The maximum number of bitcells per bitline is determined by the total charge on the bitline (precharged to V_{ref}) compared to the charge in the bitcell. The bitcell charge must be sufficient to raise or lower the bitline voltage enough to exceed the sense margin. A better sense margin allows a smaller bitline charge to be readable, and thus enables more bitcells per bitline. This is especially valuable since sense amps and decoders are used one per bitline – more cells per bitline means fewer bitlines and thus smaller chip size.
- **Smaller sense amps:** Perhaps the most obvious potential benefit. Sense amp transistors are significantly oversized versus other devices to minimize RDF. The sense amps typically take >10% of DRAM die area today – shrinking them by 50% would cut die size by 5% or more.

Summary

DRAM devices remain one of the most widely-used semiconductor devices and are critical in virtually all electronic applications. Unfortunately, continued DRAM scaling is hitting challenges in multiple areas, including reduced bitcell capacitance, scalability of the access device, scalability and cost of periphery devices, and, as discussed in this paper, continued sense amp scaling without degrading the Power, Performance, Area, and Cost (PPAC) of the overall chip.

Atomera's MST technology has a unique ability to create a quasi-delta-doped or "super-steep" channel profile to reduce mismatch and significantly improve sense amp margin, thus enabling DRAM manufacturers to lower DRAM costs and power consumption. With the emergence of AI technology, the importance of solving these problems is becoming even more critical. MST

technology provides a major step to counteract the technology headwinds faced by DRAM manufacturers giving the industry a path to lower power and higher performance solutions in the future.

References

- (1) K. Takeuchi *et al.* "Understanding Random Voltage Threshold Fluctuation by Comparing Multiple Fabs and Technologies" in *Proc. 2007 IEEE IEDM*, pp. 467-470.
- (2) N. Damrongplasit *et al.*, "Comparative Study of Uniform Versus Super-Steep Retrograde MOSFET Channel Doping and Implications for 6-T SRAM Yield," in *IEEE Transactions on Electron Devices*, vol. 60, no. 5, pp. 1790-1793, May 2013
- (3) N. Xu *et al.*, "Electron mobility enhancement in (100) oxygen-inserted silicon channel", *Appl. Phys. Lett.* 107, 123502 (2015).
- (4) J. A. Smith *et al.*, "Experimental Investigation of N-Channel Oxygen-Inserted (OI) Silicon Channel MOSFETs with High-K/Metal Gate Stack," *2018 76th Device Research Conference (DRC)*, Santa Barbara, CA, USA, 2018, pp. 1-2
- (5) H-Y Wong *et al.*, "Cryogenic Electron Mobility and Subthreshold Slope of Oxygen-Inserted (OI) Si Channel nMOSFETs" *IEEE Proc. SISPAD 2023*, September 27-28, Kobe, Japan.