Conquering the 'kink' in sub-threshold power MOSFET behavior: A simple compact modeling approach

Sachin SethArif Sonnet, January 06, 2015

Increasingly, applications such as DC/DC converters, power management units (PMU)[1], LED display power[2] and integrated display lighting solutions rely on integrated power MOSFETs (e.g., LDNMOS). Although it may be counter-intuitive, using advanced LDMOS for low-power electronic applications such as portable computing, remote sensing, implantable biomedical devices is a win-win scenario since it enables the designers to leverage both the high switching speeds, high breakdown voltage, and reduced power dissipation that these transistors offer.

To achieve the highest power conversion efficiency while also minimizing quiescent current in sleep mode (leading to increased battery lifetimes – the holy grail of power management!), these power FETs are increasingly operated in the sub-threshold or weak-inversion region of operation. This helps the design be power-frugal since the operating currents are only in the nano- to micro-ampere orders of magnitude.

**Electrical irregularities in sub-threshold**

Sub-threshold is classically defined as the range of gate-source voltages ($V_{gs}$) below the transistor's threshold voltage ($V_T$), and weak-inversion is characterized by a weak conduction channel with $V_{gs}$ barely above $V_T$. As shown in Figure 1, in these bias regions, a power MOSFET can display undesirable electrical characteristics such as humps in drain current ($I_D$) and kinks in transconductance ($g_m$), which tarnish the advantages these devices have to offer[3].
Figure 1: Normalized $I_D$ (left) and $g_m$ (right) for a wide-channel (top) and a narrow-channel (bottom) FET. The "double hump" in $I_D$ and "kink" in $g_m$ are illustrated.

Both a wide-channel FET and a narrow-width FET are shown with their corresponding transconductance $g_m$ ($dI_D/dV_{GS}$) values. All drain current values are normalized to the peak $I_D$ of the wide-channel FET.

Having such irregular $I_D$ and $g_m$ characteristics can pose immense challenges that the circuit designers have to overcome. For instance:

1. To form diff-pairs, matching the transistors with these abnormalities in subthreshold and weak-inversion regimes is exceptionally challenging and the designer is forced to operate the transistor in the strong-inversion regimes, thus increasing the overall IC power budget.
2. To get peak performance out of transistors, circuit designers bias FETs near the maximum $g_m/I_D$ dc bias point - a challenge when the dc bias that yields the maximum transconductance peak is unclear.
3. Having kinks in the $g_m$ as in Figure 1 poses immense challenges from a compact modeling perspective. A very reliable method of extracting a MOSFET's threshold voltage ($V_T$) is the $g_m$-linear-extrapolation-method[4] (GMLE), also known as the "max $g_m$" method, where $V_T$ is extracted using the following equation:

$$V_T = V_{GSmax} - \frac{I_{Dmax}}{g_{mmax}} - \frac{V_{DS}}{2}$$

In the presence of kinks, the value of $g_{mmax}$, $I_{Dmax}$ and $V_{GSmax}$ are unclear, which interferes with a clean
VT extraction of the FETs using the GMLE method. Thus, multiple design challenges are faced in the presence of irregularities in the electrical behavior of FETs in the subthreshold and weak-inversion regimes of operation.

**Cause of the "Kinks"**

As many studies have concluded, humps in $I_D$ occur due to a parasitic transistor formed during fabrication at the edges of multiple epitaxial layers and isolations[5][6][7][8]. During transistor fabrication, some acceptor dopants may seep into the oxide of the shallow trench isolation (STI) at the edges of gate (POLY) over n-type source/drain (NMOAT). This works to effectively lower the $p$-type doping under the gate region at the edges. The threshold voltage $V_T$ of a FET is a function of the doping of the back-gate:

$$V_T = V_{TO} + \left( \frac{t_{ox}}{\varepsilon_{ox}} \right) \sqrt{2q\varepsilon_S N_A (\sqrt{V_{SB} + 2\Phi_F} - \sqrt{2\Phi_F})} \quad \text{............ (1)}$$

where $V_{SB}$ is the source-to-body substrate bias, $2\Phi_F$ is the surface potential, and $V_{th}$ is threshold voltage for zero substrate bias. Due to a lower effective doping ($N_A$) at the edges (because dopants seep into STI oxide), the "edge" transistor (shown in red in Figure 2) possesses a lower threshold voltage than the "core" transistor (shown in blue in Figure 2). Traversing along the width, as we go towards the center of the FET, the doping $N_A$ is more uniform, hence the core FET remains unsullied by any edge effects.

![Figure 2: Side view (left) and Top view (right) of the cross-section of a MOSFET. The Top View shows the "core" FET (blue) and "edge" FET (red).](image)

Thus, a single FET can be imagined as being composed of two different types of FETs working in parallel. The lower $V_T$ of the edge FET makes it conduct sooner, which increases the drain current leakages in sub-threshold and weak-inversion, thereby showing the aberrant double humps in $I_D$ and kinks in $g_m$ as their electrical manifestation.

As mentioned previously, compact modeling of these kinks in power FET $d_c$ characteristics is crucial to manufacture a best-in-class PMU. Thus, to capture these electrical irregularities in a CAD-based
environment, it becomes imperative to model the edge transistor simultaneously with the core transistor. This approach is fairly robust, low cost (does not add to the model cycle time) and is fairly scalable across transistor geometries.

**Modeling the "Kink"**

The irregular double hump in $I_D$ and kink in $g_m$ can be modeled by taking two transistors connected in parallel - the core transistor and the edge transistor - in a subcircuit-like manner, where the total current is the summation of drain current contributions from each transistor:

$$I_D = q\mu_C C_{ox} \frac{W_C}{L_{eff}} (V_G - V_{TC})V_D + q\mu_E C_{ox} \frac{\Delta W}{L_{eff}} (V_G - V_{TE})V_D \quad \text{(2)}$$

Here, the suffix $E$ and $C$ stands for edge and core transistor, respectively. As the width of the core FET decreases and approaches the width magnitude of the edge FET, the current contribution of the edge transistor becomes increasingly significant. Often times, the hump in $I_D$-$V_{GS}$ characteristics indicates two different slopes in drain current $I_D$, corresponding to each of the two transistors. From Equation 2, this implies that the $\mu C_{ox}$ term for the edge transistor is different from the core transistor.

Different mobility values for the two transistors are reasonable to expect due to different residual stress in the two transistors. In such a scenario, the mobility parameters of the edge transistor can be used as a fitting parameter. Devices with double humps and kinks can be modeled by following simple prescribed steps described below:

- The oxide thickness for both core and edge FETs can initially be taken equal. Further device characterizing analysis such as Scanning/Transmission electron microscopy (SEM/TEM) could be performed to validate this assumption. For the case of SOI transistor, the oxide thickness values could potentially be very different for the two transistors.
- Determination of the edge device width is critical. For narrowest device, the impact of edge device will be most prominent. A simple linear interpolation of $I_D$ vs. Width can be used to estimate the narrow channel device width[9]. This should be a good starting point and later could be further optimized.
- Concurrent optimization of mobility, threshold voltage and sub-threshold parameters for both the core and edge transistors to capture the effect for wider width channel geometries. As discussed earlier, the edge transistor should have lower VT, smaller sub-threshold slope and lower mobility.
- Check VT and ID fitting across all widths before proceeding to next steps.
- Capture body bias dependence for both transistors. It's likely that the dependence for the two transistors is very different from each other.
- Short channel effect (SCE) should be modeled afterwards. Separate optimizations can be done to capture the SCE for both transistors.
- Retune of the width dependence can now be performed following temperature dependent parameter extraction.
- All other compact model parameters can remain the same between the core and the edge FETs.

Figure 3 shows the methodology to model irregular electrical characteristics of power MOSFETs with a fixed length (e.g., LDNMOS) in a flowchart format. The corresponding compact model parameters for each step for BSIM3, BSIM4, BSIM6, and PSP model formulations can be obtained
from their manuals.

Figure 3: A flowchart to model the kinks in $g_m$ in subthreshold and weak-inversion regimes for power FETs with a fixed length.

As a result of using this methodology, the humps in drain current and kinks in $g_m$ can be easily captured across a wide range of FET geometries, as shown in Figure 4.
Figure 4: Normalized $I_D$ (left) and $g_m$ (right) for a Wide channel (top) and a Narrow channel (bottom) FET. The "double hump" in $I_D$ and "kink" in $g_m$ are modeled correctly using a two-transistor approach.

Summary
Increasingly, savvy circuit-designers bias power MOSFETs in the sub-threshold and weak-inversion regimes of operation to maximize circuit performance and minimize IC power budgets. Doing so comes with the drawback of dealing with irregular drain current and transconductance characteristics of these transistors. Capturing these irregularities in analytical compact models is critical to enable first-pass circuit design success in power electronics, which comprises a $90$ billion market.

Footnotes:
