Solve MOSFET characteristic variation and reliability degradation issues

Takuro Nishimura - August 15, 2013

Scaling metal oxide semiconductor field effect transistors (MOSFETs) to ever smaller dimensions has delivered key benefits like performance improvement, reduced power consumption and higher-density integration. The problem is that scaled MOSFETs also face critical issues such as characteristic variation and reliability degradation.

Characteristic variation can fall into several categories like global variation, local variation, systematic variation and random variation. Among them, random variation is a strong contributor to rapid yield loss by causing a significant drop in operating margin or a malfunction of the integrated circuit, which can occur even if each stand-alone device of a circuit works properly. Random dopant fluctuation (RDF) has been a dominant factor of random variation but studies now suggest that beyond the 20-nm process node, random telegraph noise (RTN) is the main contributor of random variation. Recent research indicates that RTN is related to the bias temperature instability (BTI), which causes reliability degradation.

In scaled MOSFETs, characteristic variations caused by RTN and BTI are recognized as significant factors to device-level and circuit-level functionality, and reliability degradation, but their measurement methods are not widely understood. This article introduces the phenomena of RTN and BTI, their measurement methods, and their challenges, along with tips for handling those challenges successfully.

Random telegraph noise
RTN on MOSFET occurs when a channel carrier, a hole or an electron, is captured in an oxide trap and the captured charge is emitted from the trap. As this charge capture and charge emission continue, the drain current (Id) fluctuates, which causes the threshold voltage (Vth) to shift. The ratio of the time constant to emit a captured charge ($\tau_e$) and the time constant to capture a charge ($\tau_c$) is expressed as

$$\frac{\tau_e}{\tau_c} = A \exp\left(\frac{-E_t - E_F}{kT}\right)$$

where $A$ is the degeneracy factor, $E_t$ is the energy level of the trap, $E_F$ is the Fermi level, $k$ is Boltzmann constant, and $T$ is absolute temperature [1].

As shown in figure 1, the time-domain of an RTN signal shows binary fluctuation of Id (i.e. Vth) caused by continuous charge captures and charge emissions in a single trap. In the frequency domain, an RTN signal is inversely proportional to the square of the frequency ($1/f^2$) after a plateau in low-frequency region.
Figure 1. Time-domain (upper-left) and frequency-domain (upper-right) plots of an RTN signal show binary fluctuation of $I_d$. Histogram of $I_d$ (bottom-left) and histogram of $\tau_c$ and $\tau_e$ (bottom-right) are also shown.

Until now, the random dopant fluctuation (RDF) has been considered the dominant factor of random variation. According to a simple model of RTN, however, the amount of $V_{th}$ shift is proportional to the reciprocal of gate area (i.e. $1/LW$) for RTN [2], while it is proportional to the reciprocal square root of gate area for RDF. As a result, RTN is expected to be a main contributor of random variation at the 20-nm generation and beyond.

**Bias temperature instability**

Reliability degradation is also a difficult issue that has to be taken into account. On one hand, underestimating reliability degradation directly affects circuit functionality but on the other hand, overestimating reliability degradation makes circuit design very difficult or even impossible.

BTI is a phenomenon that introduces a significant reliability issue for the gate insulator by shifting the $V_{th}$ of a MOSFET. The degradation of $V_{th}$ due to BTI has been known since 1960s. It was recognized as a dominant reliability issue in scaled CMOS technologies in late 1990s.

Around the interface between the silicon (Si) substrate and the silicon dioxide (SiO2) gate oxide of a MOSFET, there are dangling-bond defects that build an interface state that in turn degrades the transistor characteristics. To avoid this, after the gate oxide is formed, hydrogen is annealed to bond Si-H to terminate the unattached hands of silicones.

When a negative bias is applied to the gate oxide of p-channel MOSFET, holes become majority carriers. At the Si-SiO2 interface, if a hole reacts with Si-H, it creates a dangling-bond of Si, which generates an interface state and a hydrogen ion (H+). The generated H+ is diffused into the oxide and captured. It interferes with carriers and as a result, decreases $I_d$ and shifts $V_{th}$ (see figure 2).
This phenomenon is described by reaction-diffusion (RD) theory, which had gained broad acceptance. As process nodes shrink, though, researchers observed a fast recovery phenomenon in which the degradation abated with the release of the applied stress. This prompted broad study of an RTN-like capture-emission mechanism [3]. Today, it is widely recognized that BTI consists of two components: a recoverable component, which starts recovering right after the stress is released, and a permanent component, which completely or almost completely fails to recover. It is also widely recognized that reliability degradation caused by negative BTI (NBTI) in scaled MOSFETs is more prominent than that caused by positive BTI (PBTI).

Grasser et al. indicate that RTN and the recoverable component of BTI are caused by the same defects [3], but the results reported by a different group indicate that RTN and BTI are uncorrelated threats for the device performance [4]. Finding out the whole truth about RTN and BTI will require further research with precise measurements.

**Impact to circuit design and process design**

RTN is known to increase phase noise and jitter in analog circuits and decrease the noise margin of static random access memory (SRAM) in digital circuits. The Vth shift of a MOSFET significantly affects SRAM in large-scale integrated circuits (LSI), thus RTN is a serious issue that affects the yield and performance of LSIs.

Optimizing a design margin requires enough data for statistical analysis. If the sample size is small, changes in process conditions might be hard to observe. If underestimating variability is not acceptable, statistical errors associated with the circuit simulator input must be overestimated. As process nodes shrink, designing circuits by estimating variability is getting harder and harder, and soon it will become impossible; therefore, it is important to correctly measure the variability and reasonably estimate the effect to avoid the underestimating and overestimating.

If the Vth shift caused by RTN strictly followed a simple statistical distribution, it would be possible to make a certain level of prediction from a small amount of data. Because multiple defects are involved, however, the statistical distribution is not simple and the tail of actual distribution is longer than that of normal distribution; as a result, a large amount of data is required to accurately estimate the worst value. It is thus important to appropriately predict the worst value through the statistical estimation and feed that information back to the circuit design and process design. Actually, the Vth variation of RTN for the 25-nm generation reached 70 mV [5]. The Vth variation of RTN may exceed the Vth variation of RDF as we move to future process nodes.
In case of BTI, it is necessary to assess the functionality degradations, including a fast recovery component. In terms of reliability, in order to avoid overestimating the BTI effect, it is necessary to separate a fast recovery component from other components, which is a dominant factor of reliability degradation. **Measuring RTN and BTI**

**Measuring RTN and BTI**

Now that why these measurements are important is revealed, how to make them is described next. To capture an RTN of a three-terminal MOSFET, the desired voltages between gate and source (Vgs) and between drain and source (Vds) are applied, then the drain current (Id) is measured. In order to observe variation of Id with time, Id is evaluated multiple times with a constant sampling rate.

Because the time constants of RTN highly depend on each defect of each device, they are broadly distributed from the order of 1 us or less to the order of 1 ks or more. Therefore, adjusting the sampling rate and sampling duration to the desired time range of phenomenon is important because if the sampling rate is slow, a short time constant may not be observed. If the sampling rate is fast, the amount of data may become very huge and it might be hard to handle and analyze.

There are two kinds of cycles, stress cycle and measurement cycle. When the BTI of a three-terminal MOSFET is measured, during a stress cycle, a desired Vd is applied and a constant bias (DC stress) or an arbitrary waveform bias (AC stress) is applied as Vgs. During a measurement cycle, a desired Vds is applied and a series of Id is measured with a constant Vgs (sampling (spot) measurement) or by sweeping Vgs to search the Vth (sweep measurement).

![Figure 3. NBTI AC/DC stress and Id sampling (spot)/ sweep measurement combinations](image)

The process alternates between stress cycle and measurement cycle until all stresses are applied and all measurements are done.

**Low-noise current measurement**

Capturing Id in an RTN or BTI measurement requires a current sampler or a voltage sampler such as oscilloscope or vector signal analyzer with an IV-converter consisting of resistances and amplifiers to convert current to voltage. The IV converter needs to have a low noise floor and enough bandwidth to measure the desired signal. In addition to the IV converter, voltage sources for Vgs and Vds also need to be low-noise and provide enough bandwidth to apply the desired biases.

Regarding the bias source, if an RTN measurement requires only DC bias, adding a low-pass filter
(LPF) is an option to suppress the bias noise. But since a LPF narrows the bandwidth, that method is not appropriate for a BTI measurement with AC stress or a fast BTI measurement. Thus, the bias source itself needs to be low-noise and have enough bandwidth.

Any one of the measurement units listed below is applicable for RTN and BTI measurements.
- A high-speed voltage sampler such as oscilloscope or vector signal analyzer, and IV converter. In addition to them, low-noise bias sources need to be prepared.
- A source measurement unit (SMU), which can precisely force voltage or current and simultaneously measure voltage and/or current. Bias sources are self-contained, thus no extra bias source is necessary. However, these instruments are not applicable for high-speed timing-accurate measurement such as short time constant RTN or fast component of BTI.
- Dedicated measurement apparatus designed as an all-in-one package for RTN and BTI measurements. High-speed timing-accurate measurements can be performed.

**Measurement challenges and know-how**

The time constants of RTN depend on each defect of each device under test (DUT), which means that the proper time range to observe an RTN signal is unpredictable. As a result, in order to determine an appropriate time range, sampling rate needs to be experimentally swept. For instance, it would be a good idea to measure 32,768 pts of Id with several sampling rates: 1 Msps, 100 ksps, 10 ksps and 1 ksps. If there is an RTN signal whose time constants are between about 10 μs and 1 s, they can be observed in approximately 30 s. If no RTN signal is found, a higher sampling rate (e.g. 10 Msps) or a lower sampling rate (e.g. 100 Hz) might be tried.

Note, when a high sampling rate is used, it is necessary to check whether the bandwidth of measurement apparatus is enough to measure the time range. When a low sampling rate is used, it might be necessary to restrict the bandwidth to suppress the convolution, especially when a frequency-domain analysis is performed. In the former case, all components of the measurement apparatus such as voltage sampler and IV converter need to be checked. One way to solve the latter problem is to use a higher sampling rate than the desired sampling rate and apply a digital filter (i.e., an LPF). An easier approach is to average the raw data during the sampling period, assuming the measurement apparatus has that capability. This process acts as a kind of LPF, regulating the bandwidth to suppress the convolution.

When a fast recovery component of BTI is measured, it is quite important to synchronize the stress waveform and measurement timing because the recovery starts immediately after the stress is released. The first measurement needs to start within 1 μs after the stress is released to accurately investigate the fast recovery component of BTI. If the first measurement timing is on the order of 10 ms or slower, a standard measurement unit like an SMU can perform the task but taking a first measurement within 1 μs requires a dedicated measurement apparatus.

If a recoverable component and a permanent component of BTI need to be comprehensively analyzed, the time range of BTI measurement could be 1 μs to 1 MsIt is almost impossible to measure the data with a constant sampling rate across the whole time range because the total amount of data is too huge to be handled. It would therefore be a good idea to measure the data during only a certain period of time (i.e., a measurement period) and increase a duration between one measurement period and the next (i.e., a measurement interval) logarithmically. For instance, if each measurement period is started at 1 μs, 2 μs, 5μs, 10 μs, 20 μs, 50 μs, ..., 0.5 Ms, 1 Ms, there are only 37 measurement periods and the total amount of data is feasible to be handled.

**Wafer prober**

Probe techniques can have a big effect on data quality. The following tips will help ensure accurate
RTN or BTI measurements using a wafer prober.
- The distance between a DUT and measurement apparatus should be as short as possible. A switching matrix between the two may affect noise and bandwidth adversely.
- Low-noise probers provide best results. Semi-automatic or manual probers work better than auto probers.
- Don’t pay too much attention to probe needles, power supply and room shielding—they will not significantly impact the quality of data.

When a long-time-constant RTN or long-period BTI degradation needs to be estimated, the task primarily requires very long measurement time. In this case, performing multi-device measurements in parallel is the only way to improve the performance.

To learn more about RTN, BTI and a dedicated measurement apparatus for RTN and BTI, refer [6]-[8] and [9]-[11].

Summary
In scaled MOSFETs, RTN and BTI are recognized as the most significant characteristic variation and reliability degradation issues. It is essential to understand these phenomena and their measurement methods to correctly estimate their effects on future devices and circuits. There are a lot of things to be taken into account to evaluate RTN and BTI, but once an appropriate measurement apparatus is set up and the process fine-tuned, it is easy to start exploring RTN and BTI to solve the most significant characteristic variation and reliability degradation issues.

References

About the author
Takuro Nishimura’s Profile

Also See:
- Dual regulator handles two input voltages
- How to measure MOSFET drain voltage and current

If you liked this feature, and would like to see a collection of related features delivered directly to your inbox, sign up for the Test & Measurement newsletter here.