

1st
Edition

Overcoming the Measurement Challenges of Advanced Semiconductor Technologies

DC, Pulse, and RF—From Modeling to Manufacturing



KEITHLEY

Overcoming the Measurement Challenges of Advanced Semiconductor Technologies

DC, Pulse, and RF—From Modeling to Manufacturing

1st Edition

KEITHLEY

Foreword

The editors would like to thank the Keithley employees and customers who contributed articles for the first edition of this reference work. Their contributions provided invaluable insights on the latest semiconductor technologies and the measurement solutions now emerging to address them.

Keithley contributors to the First Edition include:

Pete Hulbert, Industry Consultant

Jeff Kuo, Sr. Applications Engineer

David Rose, Sr. Staff Engineer

David Rubin, Sr. Industry Market Manager

Carl Scharrer, Principal Industry Consultant

Lee Stauffer, Lead Industry Consultant

Steve Weinzierl, Applications Engineering Manager

Yuegang Zhao, Lead Applications Engineer

Table of Contents

I Introduction

Meeting the measurement challenges of the 65nm node 2

II Advanced Transistor Gates and Thin Oxides

Integrating high frequency capacitance measurement for
monitoring process variation of equivalent oxide thickness
of ultra-thin gate dielectrics. 4

Qualifying high κ gate materials with
charge-trapping measurements 19

How to get accurate trap density measurements
using charge pumping. 29

III RF Modeling and Process Control of High Performance Analog and BiCMOS Devices

RF wafer testing: an acute need, and now practical. 38

Statistical process control of wireless device manufacturing
requires production worthy s-parameter measurements 47

IV Reliability Testing

Wafer level reliability testing—a critical device
and process development step 54

Making charge-pumping measurements with the
Model 4200-SCS Semiconductor Characterization System 62

High throughput gate dielectric reliability testing:
Digging out from the backlog 71

Improved thermal stability of copper vias using a cyclical stress test. . . . 82

Reducing parametric test costs with
faster, smarter parallel test techniques. 89

V Femtoamp DC Leakage for Mobile ICs

Tips, tricks, and traps for advanced SMU DC measurements 100

Parametric test hardware for ultra-low current measurements. 111

VI Appendix A: Selector Guides. 121

VII Appendix B: Glossary 133

OVERCOMING THE MEASUREMENT CHALLENGES
OF ADVANCED SEMICONDUCTOR TECHNOLOGIES

SECTION I

Introduction

Meeting the measurement challenges of the 65nm node

For decades, IC performance gains have largely been due to shrinking device sizes. Today, given the challenges of sub-193nm lithography, gains beyond the 90nm node are increasingly driven by material and device innovations, rather than traditional scaling. Process development engineers must leave the comfortable, well-behaved world of the Si/SiO₂/polysilicon/Al materials system and immerse themselves in the challenging world of SiGe-SOI/HfNO₂/metal gate/low κ /Cu materials.

New materials demand new electrical measurements for process and device characterization. Time-to-market pressures force researchers to choose measurement equipment suppliers who also deliver measurement expertise and complete working solutions, including subsystems from supplier partners when necessary. Increasingly, chipmakers are choosing suppliers (like Keithley) willing to accompany them throughout the long journey from research and process development to process integration and volume production.

Keithley's capability for emerging measurement needs at the 65nm node and beyond includes:

- Advanced high κ transistor gate measurements using RF C-V and charge pumping.
- Large quantities of on-wafer RF s-parameters at up to 40GHz for verifying process models.
- Isothermal DC and RF testing on SOI substrates.
- Characterizing new embedded memories like MRAM and PRAM.
- Reliability testing: NBTI, electromigration, TDDB, Cu via voiding.
- Femtoamp DC leakage on mobile ICs.
- Benchtop failure analysis.

OVERCOMING THE MEASUREMENT CHALLENGES
OF ADVANCED SEMICONDUCTOR TECHNOLOGIES

SECTION II

Advanced Transistor Gates and Thin Oxides

Integrating high frequency capacitance measurement for monitoring process variation of equivalent oxide thickness of ultra-thin gate dielectrics

Introduction

As CMOS transistors have gotten smaller and smaller, so has the thickness of their gate dielectrics. This presents a great challenge to traditional capacitance measurement used to monitor dielectric thickness for process variation. First, the relationship between the capacitance value in the inversion or accumulation region of the capacitance-voltage (C-V) curve to the gate oxide thickness is no longer simple. It's necessary to apply new models, including quantum mechanics and polysilicon depletion effects, to determine oxide thickness accurately from the C-V curve [1, 2]. Second, gate leakage increases exponentially as thickness decreases due to tunneling of carriers through the ultra-thin gate [3]. The gate capacitor becomes very lossy due to high leakage, and the gate capacitance measurement shows roll-off effects in both the inversion and accumulation regions of the C-V curve [4]. These roll-off effects make it impossible for engineers to extract C_{ox} directly and use it to monitor thickness variations in production. The roll-off behavior is also dependant on the DC leakage of the gate. Therefore, even for two gate dielectrics with the same physical thickness and area, the lower quality one with higher gate leakage will show the greater roll-off in the C-V curve, which makes it more difficult to monitor thickness variations.

Some roll-off effects in the C-V curve are device related [5, 6]. At high frequency, the two main factors are channel resistance and contact resistance. These effects could be modeled by a different equivalent circuit model and could be reduced by a new device layout. On the other hand, some of the roll-offs in C-V measurement are related to non-optimized setups, including cabling, connectors, and probe station setup [7]. The first part of the paper provides a comprehensive overview of difficulties and precautions on C-V measurement on ultra-thin gate dielectrics using LCR meters at high frequencies (1–100MHz). The second part of the paper explores C-V measurement at radio frequency (RF) as one of the approaches to solving the high leakage induced measurement problem. In general, the crossover to RFCV occurs for gate oxide equivalent oxide thickness (EOT) in the range of 1.7nm to 1.0nm.

Error analysis

Most of the challenges of using the LCR meters currently available for monitoring EOT variation come from getting correct capacitance measurements on very leaky gate materials. The effect of gate leakage in capacitance measurement can be represented by the dissipation factor (D) or quality factor (Q), where

$$D = \frac{G}{\omega C}, \text{ and}$$

$$Q = \frac{1}{D}.$$

G and C are respectively the conductance and capacitance of the gate dielectrics, and $\omega = 2\pi f$, with f being the frequency of the AC stimulus. An ideal capacitor without any parasitics has an infinite Q or zero D, while an ideal resistor has an infinite D. As gate oxide thickness decreases to less than 2nm, the effect of higher D starts to show up in the capacitance measurement. It is not unusual to see gate capacitance have D larger than 10 or even 100 at 1MHz. The direct result of large D is a roll-off of the C-V curve in the inversion or accumulation region. Sometimes, the measured capacitance value is negative [8].

Let's quickly examine how measurement error is related to D. For the simple parallel circuit in **Figure 1a**, the capacitance error can be simplified as follows:

$$\frac{\Delta C}{C} = E_0 + \Delta\theta \cdot D. \quad (1)$$

Here D can be expressed as

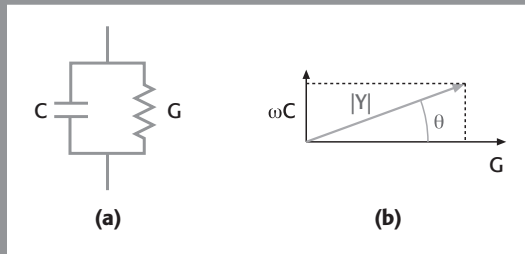
$$D + \frac{G}{\omega C} = \cot(\theta) \quad (2)$$

and E_0 denotes basic measurement error on a perfect capacitor. The definition of θ (phase angle) is shown in **Figure 1b**.

Figure 1. Schematic diagram of equivalent circuit and AC impedance measurements:

(a) Parallel equivalent circuit model

(b) Definitions of phase and amplitude of AC impedance measurements.



Eq. 1 is very important in determining errors in capacitance measurement at high D . First of all, it suggests that the measurement error is linearly dependent on D . In addition, it suggests that phase error, amplified by D , becomes the dominant source of error as D increases. There are two ways to reduce the capacitance measurement error: to increase the frequency, thereby reducing the D , or increase phase measurement accuracy, thereby reducing the phase error. Phase error comes from imperfections in the test system and measurement conditions, including cables and connectors, probes, and chuck.

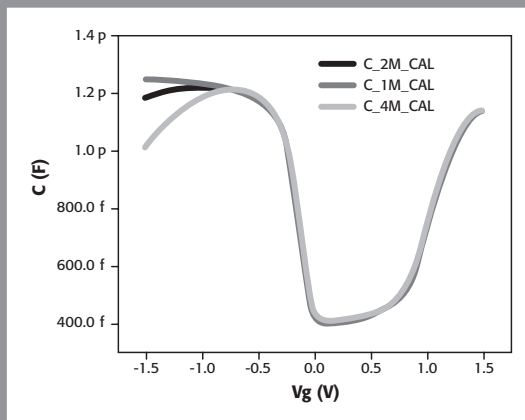
High frequency C-V measurement with LCR meters

It's not uncommon to question the minimum gate oxide thickness that current LCR meters can measure. In fact, what's important is not how thin the gate oxide is, but how leaky the gate is. As the quality of the gate oxide differs, material differs, and technology differs, gate oxides with similar equivalent oxide thicknesses may have leakage currents that differ by several orders of magnitude. One important factor to characterize the quality of the gate oxide, as described above, is D (dissipation factor) at a certain frequency (because D is inversely proportional to frequency). Since D is directly related to measured phase (θ , as shown in **Eq. 2**), the principal limitation of currently available LCR meters is their inability to resolve small phase angles due to high dissipation factor. This is mainly because the test frequency is not high enough to reduce the D factor, and there is no calibration method on those LCR meters to measure the small phase angle accurately. This sets a theoretical limit on how well those LCR meters perform on thin gate oxide measurement. On the other hand, even when using a current LCR meter, the way in which the LCR meter is set up in the measurement system also affects the quality of the C-V measurement dramatically. We will briefly review some of the improvements that can be made when setting up an LCR meter, then assess the theoretical limitation of thin oxide measurement with existing LCR meters.

Cabling is very important in C-V measurements. The overall cable length in the system must be kept as close as possible to the calibration length of the LCR meter. Any deviation in physical cable length from the calibration cable length introduces phase error. **Figure 2** shows an example of the effects of different calibration cable lengths on C-V measurements on 1.3nm gate oxide. In the measurement setup, the physical cable length is close to two meters. Different cable length values are used as inputs for cable calibration and C-V curves are measured accordingly. In **Figure 2**, we see that capacitance measurements with small D (around 0V, since hardly any DC current flows at small DC bias) are not affected very much by variations in cable length; when D is small, the overall measurement error is dominated by E_0 , according to **Eq. 1**. Phase error does not play a leading role here. On the other hand, when D is large, such as in the inversion region, where large DC current flows due to tunneling, the cable-induced phase error effect becomes significant. When the calibration length is close to the actual physical

length, the inversion region is nice and flat. However, when the calibration length is shorter than the physical length, the curve starts rolling up. When the calibration length is longer than the physical length, the curve rolls off.

Figure 2. The effect of different cable calibration lengths on C-V measurement on a 1.3nm gate oxide transistor.



Proper shield jumper location is another factor in ensuring C-V measurement accuracy. Proper operation of an LCR meter requires that the shields of the coaxial cables be properly connected as close to the DUT as possible. These shields provide a current return path that compensates for parasitic inductance from the cabling (**Figure 3a**). If the cable shields are not tied properly, close to the DUT, there will be some parts of the cables (close to the contact point to the DUT) not returning current in the shield. This results in extra phase errors due to the inductance effect. **Figure 3b** shows the effect of proper shield connections on capacitance measurement, especially when D is high. As can be seen from **Figure 3b**, the measurement without the shield jumper shows large roll-offs in both the accumulation and inversion regions, while with the shield jumper close to the DUT, there is a significant reduction in roll-offs.

For C-V measurements in production environments, probe cards for ultra-thin oxide characterization are specially designed to reduce the parasitic capacitance and inductance of wiring and probe needles. To achieve the best results, we recommend using a special probe card with minimized parasitic capacitance reserved just for thin-gate C-V measurements. When measuring a four-terminal transistor, the source, drain, and substrate are tied together in the probe card level to achieve the shortest cables possible to those terminals. Shield jumpers, which as mentioned previously are critical to capacitance measurement on leaky capacitors, are used on the probe card so that the signal path shields are tied as close to the DUT as possible.

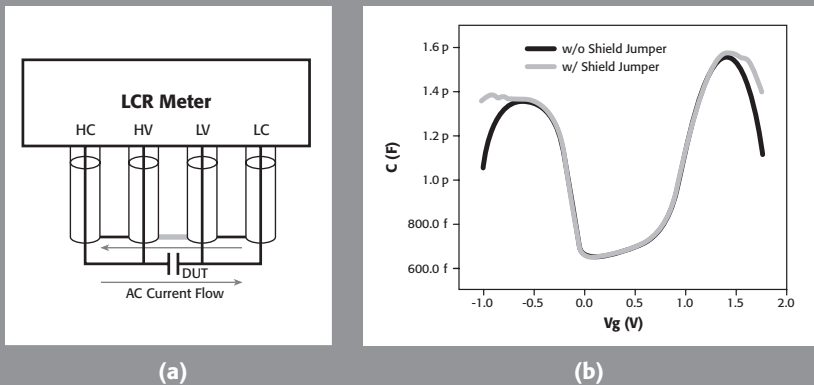


Figure 3. The effect of shield jumpers on C-V measurements:

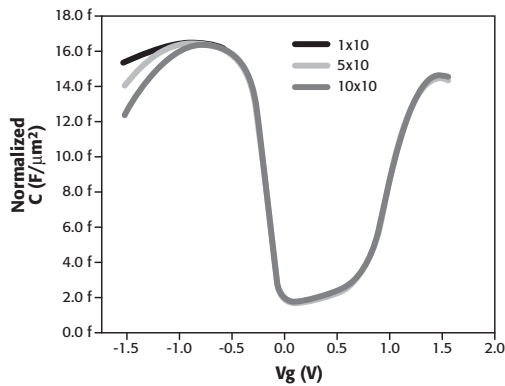
(a) Correct cabling setup for C-V measurement. The gray line between the shields of the Voltage Force and Voltage Sense terminals (which are labeled HV and LV, respectively) is the shield jumper that ties the shields of the cables together close to the DUT.

(b) The effect of the shield jumper on the C-V measurement of a 1.3nm gate oxide. The C-V curve shows less roll-off in the inversion region with a properly located shield jumper.

Cable calibration is critical to successful C-V measurement on high D capacitors. Cable calibration includes open, short, and load calibrations. While both short and load calibrations require a proper test structure on the wafer (e.g., short and 50Ω load), open calibration does not. It has been found that the quality of short calibration determines the overall measurement quality. When calibrating on a short structure, there are inevitably some contact resistances. Short calibration with a high contact resistance results in noisy measurements and roll-offs in the C-V curve. Therefore, it is crucial to reduce contact resistance as much as possible during calibration. In a production environment, it is crucial to have an auto-calibration procedure. The system can be set up so that it performs calibration automatically when certain calibration criteria are met. Those criteria include the duration of the previous calibration, whether the previous calibration failed or not, or whether it is the first time to calibrate. To be consistent, calibration is only performed on the first wafer of every cassette. The user can change calibration criteria according to specific needs.

It is well known that some roll-offs in C-V curves are due to the channel resistance of the MOSFET [6]. **Figure 4** shows an example of capacitance measurements on three transistors with different gate lengths with area normalized capacitance value. Those transistors are on the same site on a wafer and very close to each other. It clearly shows that channel resistance-induced roll-off effect. Even though this effect can be compen-

Figure 4. The effect of channel resistance on C-V measurement on 1.3nm gate oxide. Three curves represent C-V measurements of $1\mu\text{m}\times 10\mu\text{m}$, $5\mu\text{m}\times 10\mu\text{m}$, and $10\mu\text{m}\times 10\mu\text{m}$ transistors respectively. Measurement on shorter channel length shows less roll-off in the inversion region.

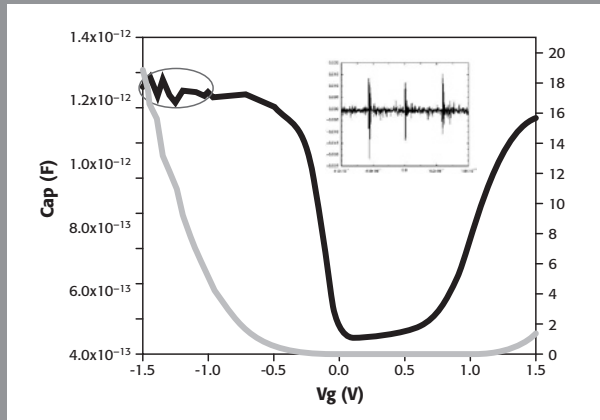


sated for by proper device modeling, it's undesirable in a production environment. We recommend using a short channel transistor to minimize channel resistance-induced roll-off. The trade-off of using a small area transistor is that the fringing capacitance is relatively large. Fortunately the fringing capacitance can be subtracted by measuring capacitances with two different gate areas. The area of the gate should be designed so that the capacitance value to be measured is around 1–2pF. Higher capacitance values result in measurement range overload due to high leakage, while lower capacitance values result in noisy measurements due to resolution limitations of the LCR meter.

Besides measurement accuracy, noise is another important factor with thin gate oxide C-V measurement. Again, based on **Eq. 1**, measurement noise is directly proportional to D factor. As D factor increases dramatically as gate thickness decreases, measurement noise, or repeatability of the measurement, which was not a problem before, becomes an issue. The most common source of noise is the chuck, especially a thermal chuck. If the device under test is not isolated from the chuck, as is typically the case with an NMOS transistor in a CMOS process, chuck noise can couple into measurement and becomes obvious when D is high. **Figure 5** shows the effect of measurement noise coupled with noise from a thermal chuck. There are several solutions to this problem:

- Use a better isolated, low noise chuck.
- Use higher frequencies so that D is reduced.
- Turn off the power to the thermal chuck when it's not in use.
- Design the test structure so that the chuck is isolated from the body of the transistor.

Figure 5. Example of chuck noise coupled into C-V measurement noise at high D. The insert in the graph shows scope plots of the chuck noise.



One common misunderstanding is that the higher the test frequency used, the better the capacitance measurement on thin gate oxide will be. In principle, this is true, because D is smaller at higher frequencies (this is usually true for frequencies less than 100MHz, where the DUT can be represented by a parallel equivalent circuit). However, to implement this principle with the LCR meters currently on the market, one other important factor must be considered, which is their basic measurement accuracy at higher frequencies. This is related to the E_0 term in **Eq. 1**. E_0 represents the measurement accuracy on an ideal capacitor ($D = 0$). E_0 is a function of frequency. By reviewing the published specifications for currently available LCR meters [10], one can easily learn that the measurement accuracy degrades as frequency moves toward the high end of the frequency range (1–100MHz).

Another way to look at the problem is to combine dissipation factor, frequency, and the measurement specification for the LCR meter and draw a plot of measurement accuracy as a function of frequency. **Figure 6** shows that on a gate dielectric with a $D = 10$ at 1MHz, the least error occurs at frequencies of around 1MHz. Therefore, a LCR meter with 1MHz capability should be sufficient for measuring gates with $D = 10$. For example, as shown in **Figure 7** for a 1MHz C-V measurement on a 1.3nm oxide—the largest D at that frequency is close to 20.

Following a similar approach, if a gate dielectric has a $D = 100$ at 1MHz, the sweet spot moves to a higher frequency (100MHz). However, at that frequency, the lowest measurement error rises to around 10%, which may be unacceptably high. To go even further, for a gate with a $D = 1000$ at 1MHz, which is equivalent to $D = 10$ at 100MHz, the minimum measurement error for frequencies up to 100MHz is around 30%, which

Figure 6.
Example of
specification
error of an LCR
meter for $D = 10$
(1MHz) across
frequency.

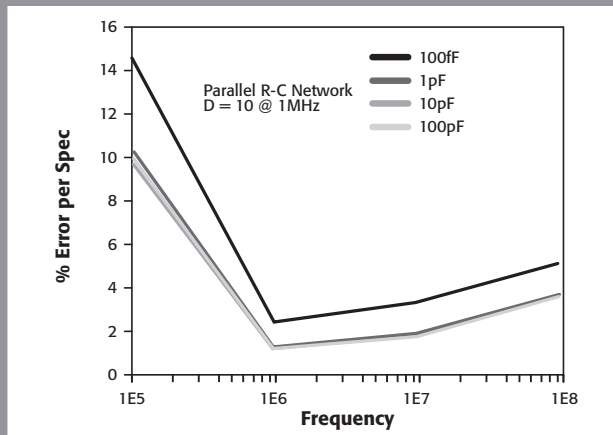
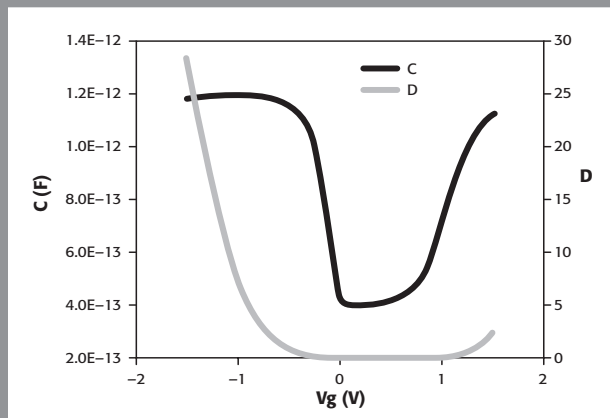


Figure 7.
Example of C-V
measurement at
1MHz on 1.3nm
gate oxide.



makes this LCR meter completely unsuitable for this type of measurement. This exercise sets a theoretical limit on the maximum leakiness of an oxide that can be measured using the currently available LCR meters that operate at frequencies from 1 to 100MHz. Depending on the D of the material being measured, this exercise may be useful in determining the level of leakage at which a specific LCR can no longer measure a particular oxide. Alternatively, it's also possible, for a given D , to determine the optimal measurement frequency.

RF capacitance measurement

With current LCR meters, the product of phase measurement accuracy and D, as in **Eq. 1**, is limited, which limits the capacitance measurement accuracy. At 110MHz, the infrastructure, including cabling, probe card, and calibration, is similar to that of an RF measurement. It requires a full calibration set, including phase, open, short, and load calibration. At the same time, performance is again limited by the product of phase measurement accuracy and D. With an alternative approach, using the RF technique, one can measure capacitance at much higher frequency, in particular at operating frequency, such as 2.4GHz. Usually the measurement is done at a frequency greater than 1GHz, the point of maximum Q. At such frequencies, D will remain relatively small for the foreseeable future (according to the International Technology Roadmap for Semiconductors) [11]. Conductance due to leakage ceases to be an issue.

The RF capacitance of the DUT is derived from the complex conductance (Y)

$$C = \frac{|Y|^2}{2\pi f \operatorname{Im}(Y)},$$

which is calculated from s-parameters measured on a two-port network (**Figure 8**). A vector network analyzer is used to measure the RF scattering parameters. The characteristic impedance of the overall transmission line of the system is optimized for 50Ω, with a 20GHz bandwidth. RF signals from the VNA are passed to the RF probe card through a dedicated pathway. The DC bias and RF signal are mixed in the Keithley S600 Series testhead in very close proximity to the DUT. The component of complex impedance (Z) of the DUT can be calculated from reflection parameters S_{11} and S_{22} ,

$$S_{11} = \frac{Z_{11} - Z_0}{Z_{11} + Z_0} \quad \text{and} \quad S_{22} = \frac{Z_{22} - Z_0}{Z_{22} + Z_0}$$

with $Z_0 = 50\Omega$. The frequency of measurement is selected so that the AC impedance of the DUT is close to 50Ω, because the measurement accuracy of the vector network

Figure 8. Schematic of a two-port network for s-parameter measurement.

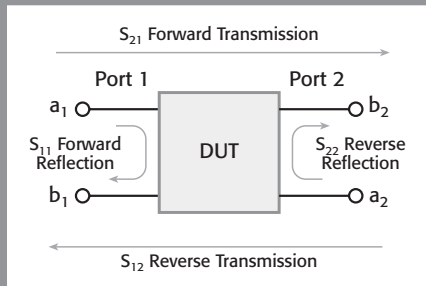
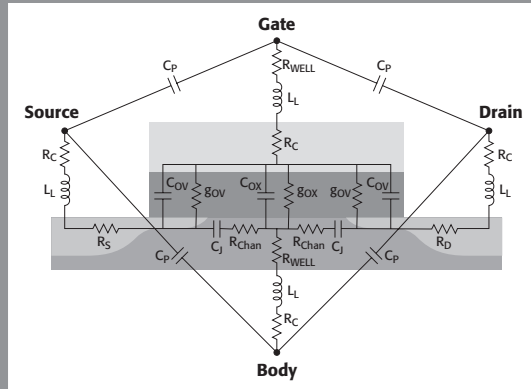


Figure 9. Simplified circuit model of a MOSFET including imperfections. The main factors to consider are parasitic capacitance between contact pads and leads (C_p), contact resistance (R_c), lead inductance (L_l), channel resistance (R_{ch}), and overlap capacitance (C_{ov}). Most of the imperfection factors can be corrected by de-embedding.



analyzer is optimized around 50Ω impedance. For example, a 1pF capacitor has 50Ω impedance at around 3GHz .

The parasitics embedded in the measurement system and the DUT are part of the technical difficulties involved in using RF measurement. Significant work has been done in measurement methodology and device layout to de-embed the parasitics. **Figure 9** shows a simplified circuit model of a real transistor. The goal of RF capacitance measurement is to get C_{ox} . However, C_{ox} is surrounded by imperfections in the physical device. Those imperfections include overlap capacitance between the gate contact and the source/drain well, gate resistance (due to poly silicon), lead inductance (from DUT to contact pads), contact resistance (between probe needle and contact pads), and channel resistance (mentioned previously). Some of the imperfections can be extracted by the de-embedding technique, especially the effects of contact resistance, lead inductance, and parasitic capacitance. DUT measurement results are corrected by subtracting those measured on de-embedding structures. Typical de-embedding structures include open, short, and thru (**Figure 10**). For short de-embedding,

$$Z = Z_{meas} - Z_{short} ,$$

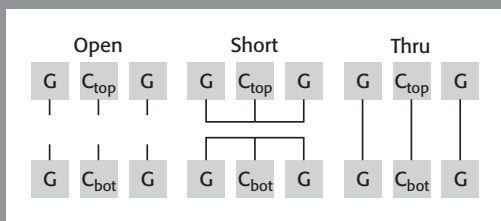
for open de-embedding,

$$Y = Y_{meas} - Y_{open} ,$$

and for thru de-embedding,

$$\frac{1}{Y} = \frac{1}{Y_{meas}} - \frac{1}{Y_{thru}} .$$

Figure 10. De-embedding structures layouts.



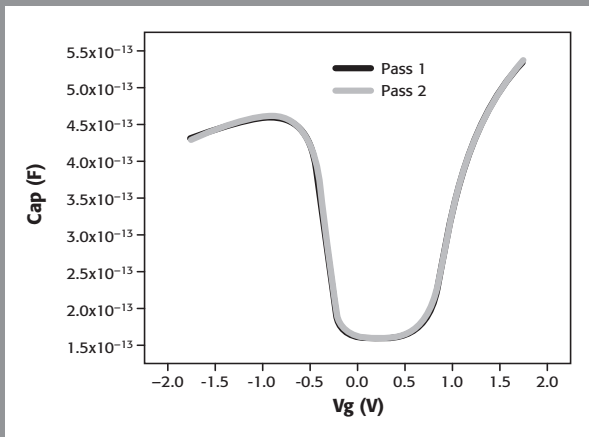
Combinations of two or more of the de-embeddings can also be used. It is common for open and short de-embeddings to be used together to correct both parasitic capacitance and contact resistance. Sometimes open and thru are used together when the series inductance of the DUT is not optimized.

One basic assumption of RF C-V measurement is that the characteristic impedance of the system's transmission line is 50Ω . The closer the impedance of the transmission line to 50Ω , the better the measurement result will be. The device layout on the wafer should be adjusted to match the transmission line impedance. A ground-signal-ground structure is required for RF measurement. As mentioned earlier, channel resistance effects will show up in C-V measurements on long-channel devices, especially at higher frequencies. It is recommended that small transistors be used for RF capacitance measurements. To achieve a better signal-to-noise ratio, many small area transistors can be connected in parallel to make a large device. More details on the design of test structures for RF C-V measurement are available [9].

Contact resistance variations between consecutive probe contacts can limit the repeatability of RF C-V measurements. For example, if the DUT has a characteristic impedance of $100k\Omega$ at 1MHz, a 1Ω variation in contact resistance will cause only a 0.01% error. However, the characteristic impedance of the same device drops to 100Ω at 1GHz, so that same 1Ω variation in contact resistance will induce a 1% error. Most of the variation in contact resistance is due to buildup of aluminum oxide on the tip of the probe needle. These variations have been engineered out of the S600 Series through the use of automated probe cleaning. A Gage R&R study shows less than 5% variation in most cases (30% variation in Gage R&R is considered "good"). **Figure 11** shows an example of repeated RF C-V measurements.

Calibrations down to the probe tips are required to make accurate measurements. A full calibration set includes open, short, thru, and load calibrations. Calibration will compensate for imperfections in the transmission line, including parasitic capacitance and lead inductance on the probe card and connectors. However, calibration cannot compensate for contact resistance, because the contact resistance between the probe

Figure 11. Overlay of two RF C-V measurements at 2.4GHz.



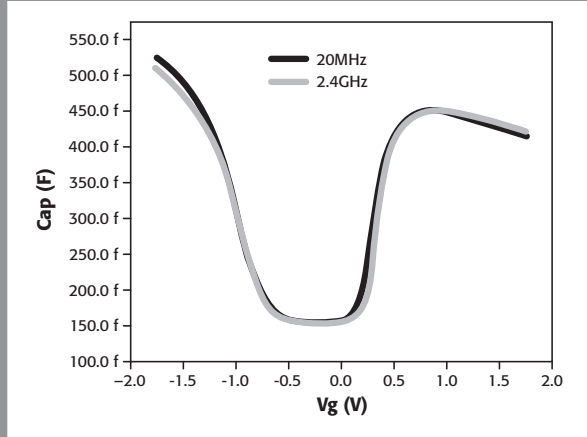
needle and the gold contacts on the calibration substrate is not the same as that between the probe needle and the aluminum pad on actual wafer under tests. These subtle differences are also compensated for in the S600 Series automation.

A full suite of RF capacitance measurements involves:

- Loading calibration wafer and performing calibration. This is required only if the probe card is changed or if more than 72 hours has elapsed since the last calibration. The S600 Series accomplishes this in a way that is compliant with all 300mm automation requirements.
- Loading wafers from cassette.
- Performing s-parameter measurements on de-embedding structure (once per lot).
- Performing s-parameter measurements on actual DUT. The resolution of the Keithley system is sufficient to measure a single 100fF DUT, extract gate and fringing capacitance, and correct for poly depletion.
- Outputting de-embedded results.

Figure 12 demonstrates the correlation of RF C-V measurement results on a 13Å gate oxide at 2.4GHz with a 20MHz C-V measurement using an LCR meter. It shows excellent agreement between the two methods.

Figure 12. RF C-V measurement on a 1.3nm gate oxide.



Comparison of high frequency and RF C-V techniques

Comparing high frequency C-V (HFCV) and radio frequency C-V (RFCV) results only makes sense when both techniques can yield valid EOT measurements. In general, this applies to gate oxides with EOTs typically ranging from 1.7nm to 1.0nm. The actual range varies, depending on the specific process. For thick oxide, HFCV has a clear advantage because of cost and ease of use. For ultra-thin oxide, HFCV is no longer capable because of the reasons stated previously in this paper. When both HFCV and RFCV are valid options, the comparison shown in **Table 1** may help users determine the best time to migrate to RFCV based on cost, technology roadmap, and other factors.

Table 1.

Factors	Technique	Implementation	Advantage	Disadvantage
Device layout	HFCV	DC	Compatible with existing DC parametric tests	Cabling and connection becomes harder at higher frequency ¹
	RFCV	G-S-G, RF De-embedding	Parasitic extraction	Not compatible with DC parametric tests
Device size	HFCV	Large	Less parasitic compared to gate capacitance	Channel resistance affects C-V measurement
	RFCV	Small	Can measure short channel devices, reducing channel resistance effect Can measure working transistors, not test structures	Parasitic capacitance has to be extracted with de-embedding
Contact resistance	HFCV	Dual frequency C-V sweep		Requires two sweeps Accuracy is limited by instrument accuracy ²
	RFCV	De-embedding	Only one frequency is needed	
Frequency	HFCV	100kHz – 100MHz		
	RFCV	>100MHz	Measurement at operating frequency	
DC current	HFCV	DC current flow into meter		Measurement accuracy affected by DC current ³
	RFCV	DC current is separated from RF pathway ⁴	RF measurement not affected by amount of DC current flow	

Notes:

1. At frequency >1MHz, special care has to be taken for cabling and connection (such as a dedicated signal pathway), an RF-like calibration suite, such as open, short, load, has to be deployed to calibrate down to the probe tip.
2. See [12].
3. See [8].
4. DC bias is provided by a source measure unit through a bias Tee, while RF measurement is AC coupled through the bias Tee using a Vector Network Analyzer.

Conclusion

Common C-V measurement errors with currently available LCR meters are discussed, as well as their limitations in making C-V measurements on ultra-thin gate oxides. Techniques to enhance an LCR meter's performance to near its theoretical limits, such as cabling, probe card, and connectors, are discussed. The RF C-V technique is discussed and deployed in production environment to monitor EOT variations for ultra-thin gate oxide. New options available make the S600 Series the ultimate tool for monitoring EOT variation in production environments for the current technology node, as well as for several future technology nodes. ■

References

- [1] N. Yang, W. K. Henson, J. R. Hauser, and J. J. Wortman, "Modeling Study of Ultrathin Gate Oxides Using Direct Tunneling Current and Capacitance-Voltage Measurements in MOS Devices," *IEEE Transactions on Electron Devices*, vol. 46, p. 1464, July 1999.
- [2] K.F. Schuegraf, et al., "Impact of Polysilicon Depletion in Thin Oxide MOS Technology," in *Proc. VLSI-TSA*, 1993, p. 86.
- [3] Y. Shi, T. P. Ma, S. Prasad, and S. Dhanda, "Polarity-dependent tunneling current and oxide breakdown in dual-gate CMOSFETs," *Electronics Device Lett.*, vol. 19, p. 391, Oct. 1998.
- [4] C. H. Choi, et al., "Capacitance reconstruction from measured C-V in high leakage, nitride/oxide MOS," *IEEE Transactions on Electron Devices*, vol. 47, p. 1843, Oct. 2000.
- [5] K. J. Yang and C.M. Hu, "MOS Capacitance Measurements for High-Leakage Thin Dielectrics," *IEEE Transactions on Electron Devices*, vol. 46, p. 1500, July 1999.
- [6] D. W. Barlage, et al., "Inversion MOS Capacitance Extraction for High-Leakage Dielectrics Using a Transmission Line Equivalent Circuit," *IEEE Electron Device Letters*, vol. 21, p. 454, Sept. 2000.
- [7] H. Suto, et al., "Methodology for Accurate C-V Measurement of Gate Insulators below 1.5nm EOT," in *Extended Abstract of the International Conf. On Solid State Devices and Materials*, 2002, p. 748.
- [8] Y. Okawa, H. Norimatsu, H. Suto, and M. Takayanagi, "The Negative Capacitance Effect on the C-V measurement of Ultra Thin Gate Dielectrics Induced by the Stray Capacitance of the Measurement System," in *Proc. ICMTS*, 2003, p. 197.
- [9] J. Schmitz, et al., "Test Structure Design Considerations for RF-CV Measurements on Leaky Dielectrics," in *Proc. ICMTS*, 2003, p. 181.
- [10] Agilent 4294 LCR meter specification, document #5968-3809E.
- [11] ITRS website: <http://public.itrs.net/>.
- [12] A. Nara, N. Yasuda, H. Satake, and A. Toriumi, "Limitations of the Two-frequency Capacitance Measurement Technique Applied to Ultra-Thin SiO₂ Gate Oxides," in *Proc. ICMTS*, 2001, p. 53.

Qualifying high κ gate materials with charge-trapping measurements

Hunting for high κ

As the size of transistors continues to scale down, the use of conventional SiO_2 as a gate dielectric material is approaching physical and electrical limits [1, 2]. The principal limitation is high leakage current due to quantum mechanical tunneling of carriers through the thin gate oxide [3]. To reduce gate leakage current, high dielectric constant (high κ) gate materials, such as HfO_2 , ZrO_2 and Al_2O_3 and their silicates [4], have drawn a great deal of attention in recent years. Due to their high dielectric constants, high κ gates can be made much thicker than SiO_2 while achieving the same gate capacitance. The result is lower leakage current—sometimes, several orders of magnitude lower.

One of the remaining challenges of deploying high κ materials is reliability. This includes phenomena affecting material reliability, such as voltage breakdown and defect generation mechanisms, and phenomena affecting device reliability, such as hot carrier injection. To characterize the reliability of high κ gate materials fully, multiple measurement techniques are typically required. Usually, these techniques include I-V, C-V, charge-pumping, and other measurements.

Various instruments can be used to take these measurements, but a fully integrated device characterization test system speeds up testing and provides a high level of data integrity. These systems typically integrate source-measure units with a C-V meter and pulse generator to characterize charge-trapping phenomena inside the high κ gate material. They can be used with various charge-trapping measurements, including a relatively new stress and charge-pumping technique that better characterizes traps in high κ films.

Overview of charge-trapping measurement techniques

Charge-trapping techniques involve a series of voltage stresses of certain duration. During voltage stress, leakage current is measured in real time to calculate the amount of charge injected into the gate. This quantity is expressed as:

$$Q_{inj} = \int I_{Leakage} / dt.$$

Between voltage stresses, three types of measurements can be done in sequence: C-V, I-V, and charge-pumping. From these measurements, important device parameters can be extracted and plotted as a function of time to show the degradation caused by the stresses.

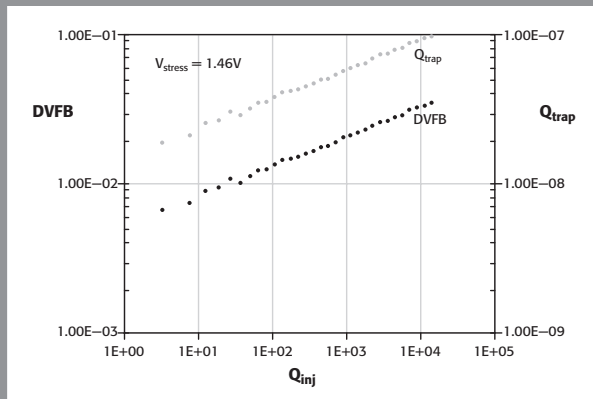
Stress and C-V Measurements [5]. In this measurement, the device under test (DUT) usually is a MOS capacitor. A C-V sweep is performed on the DUT before and after voltage stress. The C-V sweep can be a full sweep from inversion to accumulation, so that a flat band voltage can be calculated by quantum mechanical modeling. However, a faster and easier way is to do the voltage sweep in a relatively small voltage range around an estimated flat band. The flat band voltage is then extracted from the C-V data. Either a single sweep or bi-directional sweep can be used; a bi-directional sweep will show any hysteresis effects. Flat band voltage as a function of stress time or injected charge provides information on how much charge is trapped in the gate stack structure. The trapped charge may be characterized by an effective value assumed to be located at the insulator-silicon interface (Q_{trap}), given by:

$$Q_{trap} = C_{gate} \cdot \Delta C_{fb}.$$

Trapped charge calculated from the change in flat band voltage is an approximation of the charge generated in the semiconductor-insulator interface. However, since charge can be generated in places other than this interface, stress and C-V measurements only give a rough estimate of how much is trapped due to injected charge. (See sample data in **Figure 1**.)

This measurement technique has the advantage of being simple and direct. It measures the effect of trapped charges from the C-V curve shift along the voltage axis as a function of injected charges. However, it is essential to avoid relaxation of trapped charges during the stress cycle. If trapped charges de-trap too fast, some of the trapped charges may be lost during switching between the stress and C-V measurements. Minimizing the switching time is the key to success in this measurement. Another drawback

Figure 1. A sample plot of trapped charges and flat band voltage shifts vs. injected charges for HfO_2 film with Equivalent Oxide Thickness = 1.2nm.



of this technique is that it measures the combination of traps initially in the film, plus those created later by the stress.

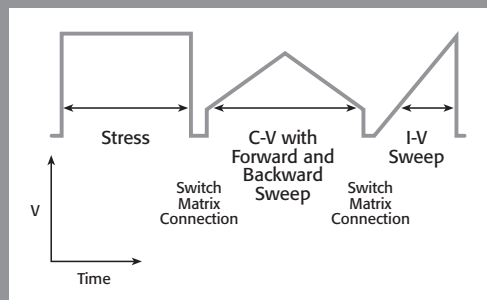
Stress and I-V Measurement. Another method similar to the stress and C-V measurement is stress and I-V measurement on a MOSFET. During stress, the source, drain, and substrate terminals are grounded, so that stress is only applied on the gate dielectric. Then, after stress, a V_{gs} - I_d test is performed, so that key parameters, such as threshold voltage and channel transconductance, are extracted. Plotting the shift of those parameters as a function of injected charge makes it possible to obtain the trapped charge density.

This method requires only I-V measurements, so it offers the advantage of being conducted without the need for a switching matrix. This can avoid or significantly reduce charge relaxation effects. However, modeling work is required to interpret the data.

Figure 2 is an example of voltage waveforms applied to an MIS capacitor in one stress cycle that includes stress, C-V, and I-V measurements. A forward and backward C-V sweep (center waveform in **Figure 2**) might be used to uncover any hysteresis effect in the high κ dielectric film. The instrumentation must be switched from stress to C-V measurement, so a voltage discontinuity appears at the DUT terminals during the switching time. This voltage discontinuity could result in relaxation of trapped charges from trapping centers. If so, the C-V measurement afterwards would indicate a smaller flat-band voltage shift due to fewer trapped charges. Therefore, the switching time between instruments must be minimized.

Charge-Pumping Measurement. Charge-pumping measurements are widely used to characterize interface state densities in MOSFET devices. This type of measurement is especially useful for thin gate materials that have relatively large gate leakage currents when accurate removal of the gate leakage is done [6, 7]. Such leakage makes it difficult, if not impossible, to collect simultaneous quasistatic and high frequency C-V measure-

Figure 2. Applied voltage waveform diagram for stress and C-V/I-V measurement.



ment data needed to estimate interface state densities. The interfacial-trapped charge (D_{it}) is calculated by:

$$D_{it} = \frac{I_{cp}}{qAf\Delta E},$$

where I_{cp} is the measured charge-pumping current, q is the fundamental electronic charge, A is the area, f is the frequency, and ΔE is the difference between the inversion Fermi level and the accumulation Fermi level [8].

The basic charge-pumping technique involves the measurement of the substrate current while applying voltage pulses of fixed amplitude, rise time, fall time, and frequency to the gate of the transistor with the source, drain, and body tied to ground (**Figure 3a**). The application of the pulse can be done with a fixed amplitude voltage base sweep or with a fixed base variable amplitude sweep.

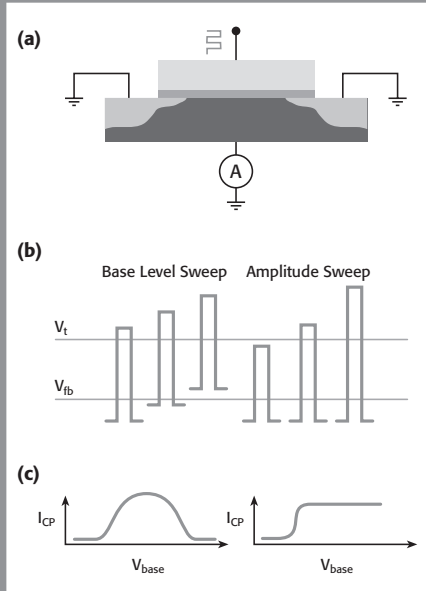
In a voltage base sweep, the amplitude and period (width) of the pulse are fixed while sweeping the pulse base voltage (**Figure 3b**). At each base voltage, body current can be measured and plotted against base voltage (I_{CP} vs. V_{base}). The interface trap density (D_{it}) as a function of band bending can then be extracted from the charge-pumping current if the ΔE is known.

Figure 3. Overview of charge-pumping measurements:

(a) Schematic for charge-pumping measurement; source and drain of the transistor are connected to ground; gate is pulsed with fixed frequency and amplitude while body current is measured.

(b) Pulse waveform for base voltage sweep; pulse amplitude is constant.

(c) Pulse waveform for amplitude sweep; base voltage is constant.



A fixed base, variable amplitude sweep has a fixed base voltage and pulse frequency with step changes in voltage amplitude (**Figure 3c**). The information obtained is similar to that extracted from a voltage base sweep, but in this case, I_{CP} vs. V_{peak} is plotted. These measurements can also be performed at different frequencies, so that a frequency response of interface traps can be obtained.

For high κ gate stack structures, the CP technique can quantify the trapped charge (N_{it}) as:

$$N_{it} = \frac{I_{cp}}{qfA},$$

because trapped charge beyond the silicon substrate/interfacial layer can be sensed [9]. **Figure 4a** shows the characteristic N_{it} curve for the base voltage sweep technique, while **4b** shows the N_{it} characteristic for the fixed based, variable amplitude technique.

Stress and Charge-Pumping Measurement. Stress C-V, stress I-V, and charge pumping can provide information about charge centers associated with defects already in a gate dielectric film. However, for stress C-V and stress I-V, it isn't possible to distinguish between charge centers initially in the film and those created during a measurement stress cycle. Both types contribute to the shift in flat band voltage during measurements.

However, a recently developed technique can distinguish the initial charge-trapping centers from those created later in the film by voltage stresses. This technique uses a combination of stress and charge-pumping measurements. A major advantage of the new technique is that relaxation of trapped charges during the stress cycle will not affect overall measurement accuracy. The charge-pumping measurements detect traps in high κ gate stacks, so with some modeling work, it is possible to compare results of charge densities before and after a stress cycle. This indicates how many new charge centers were created by injected charges.

Figure 5 shows results from stress charge-pumping measurements on a nMOSFET with $W/L = 10/1\mu\text{m}$. The gate stack is an ALD HfO_2 with a chemically grown interfacial oxide (EOT) of 1.7nm [10].

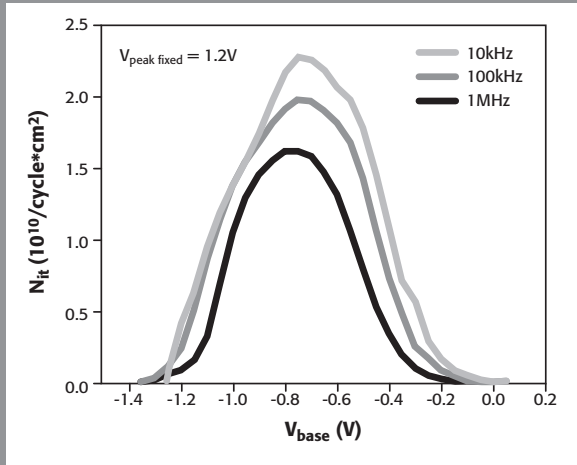
Test hardware arrangement

The core instrumentation for these measurements is a semiconductor characterization system (SCS) with multiple source-measure units (SMUs) and pre-amps that provide sub-femtoamp resolution for gate leakage currents. This instrumentation is combined with a capacitance measuring instrument, pulse generator, and semiconductor switching matrix for a complete measurement solution. An example of such a system appears in **Figure 6**. The equipment in this block diagram includes a:

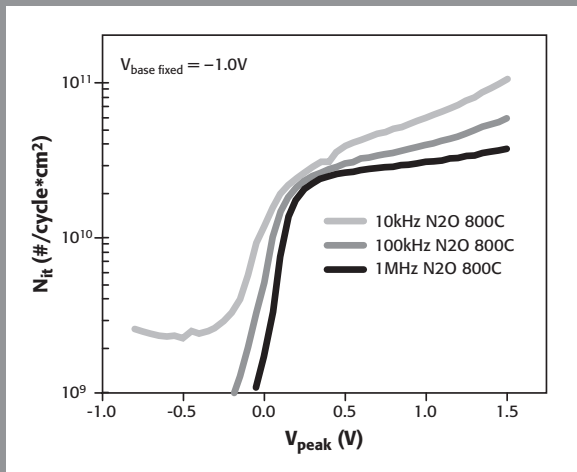
- Keithley Model 4200-SCS Semiconductor Characterization System

Figure 4. Examples of charge-pumping measurements on MOSFET with high κ gate materials.

(a) Conventional base sweep at different frequencies.



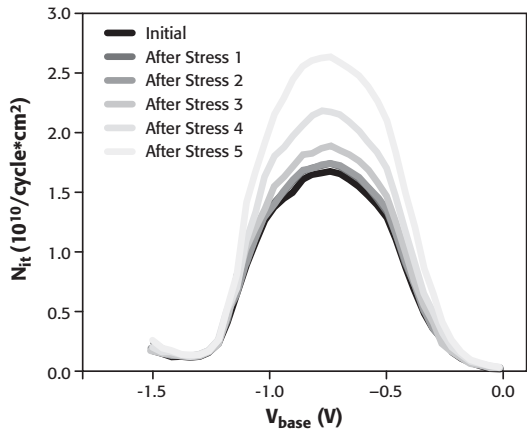
(b) Amplitude sweep at different frequencies.



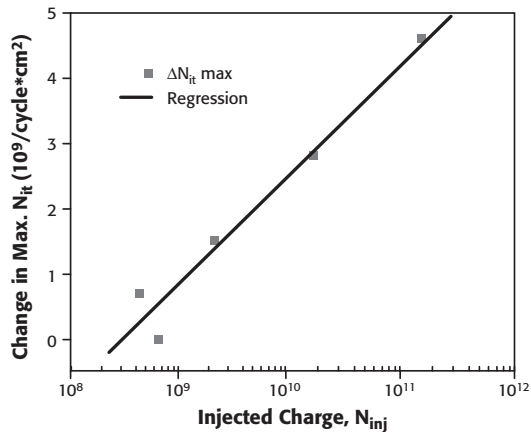
- Keithley Model 590 C-V Meter or Agilent Model 4284 LCR Meter
- Agilent 8112 or 8110/81110 Pulse Generator Unit (PGU)
- Keithley Models 707A and 708A switching mainframes with Model 7174A ultra low leakage switch matrix cards

Figure 5. Stress and charge-pumping measurement on a nMOSFET transistor with an ALD HfO_2 gate dielectric and chemical oxide interfacial layer.

(a). Result of the stress with interspersed CP measurement after stress.

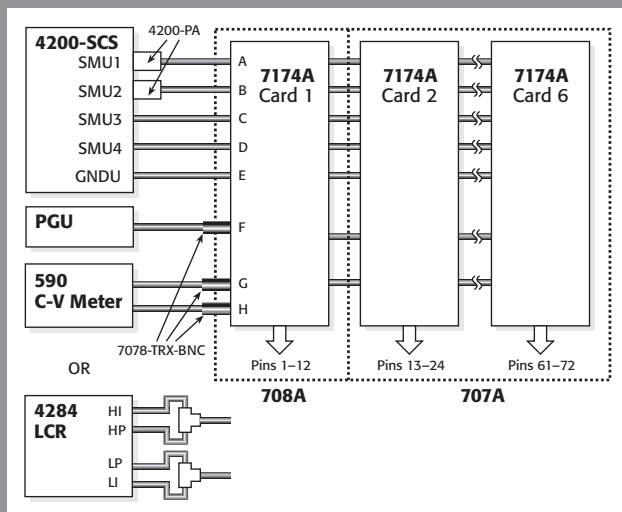


(b) Increase of maximum interface trap density as a function of number of injected charges.



The pulse generator supplies the voltage pulses for charge-pumping measurements. The pulse generator and the C-V meter are connected to the rows of the switch matrix card. DUTs are connected to the columns of the matrix cards. The low leakage and minimal dielectric absorption of the 7174A cards ensure that DUT measurements can be made much faster and more accurately than with general purpose switching cards. The 4200-SCS also has dynamic Trigger Link outputs for control of internal and external in-

Figure 6.
Semiconductor
characterization
system diagram.



strumentation without using the GPIB, which also speeds up measurements. Its probe drivers provide manual and automatic control of on-wafer measurements.

System software and data communications

For this hardware, the test application is written for use with the Microsoft® Windows® operating system running on the PC in the SCS. The software provides test definition, automated control, parameter analysis, and data graphing. Built-in measurement configurations include a sweep mode with nine forcing functions, which reduces programming requirements. Communications between the CPU mainboard and SMUs takes place over a PCI interface, which is much faster than a GPIB interface. This is a key feature in minimizing switching time between stress, C-V, and I-V measurements.

Figure 7 shows two screen captures of the charge-trapping software interface for a test setup and a charge-pumping data plot. To get a complete picture of charge-trapping phenomena in high κ dielectric materials, it's necessary to configure individual tests for all the plausible combinations of stress, C-V, I-V, and charge-pumping measurements. Typical test variables are listed in **Table 1**. Depending on the SCS, data could be stored in text or Excel format for post-processing.

Acknowledgements

The charge-trapping test application developed with the Keithley Model 4200-SCS was the result of a collaborative effort between Keithley, International SEMATECH

Figure 7. Charge-trapping software interface.

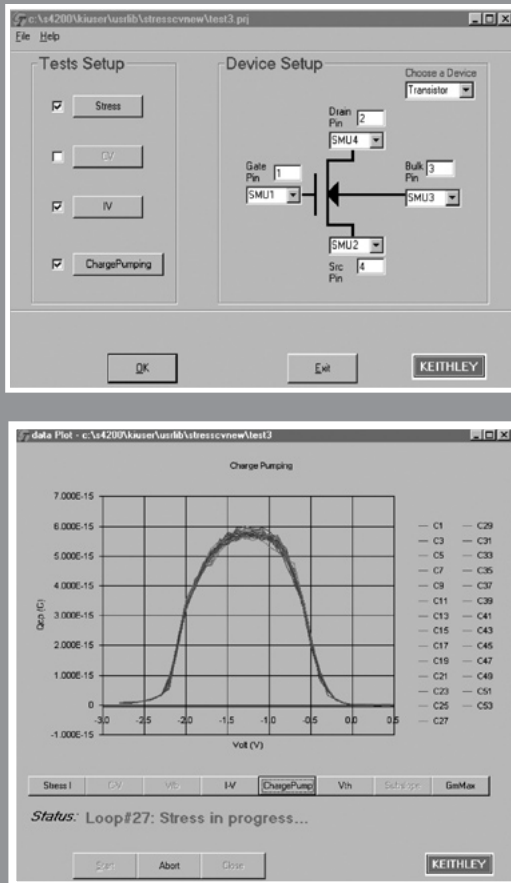


Table 1. List of tests and parameters extracted on charge-trapping test system

Test Type	Stress-Only	Charge-Pumping Only	Stress – C-V/I-V	Stress – Charge Pumping
Measurement Performed	Constant Voltage Stress or Voltage-Ramp	Base Voltage Sweep Amplitude Sweep Frequency Sweep	Dual C-V Sweep V_g - I_d Test	Stress and Charge-pumping
Parameters Extracted	TDDB or Q_{BD} *	I_{cp} , Q_{cp} , D_{it}	Shift of V_{fb} , V_{th} , G_m , sub-threshold slope as function of injected charges	Charge Density, new charge created due to injected charge

* TDDB—Time Dependent Dielectric Breakdown; Q_{BD} —Charge to Breakdown.

(ISMT), and IMEC. Special thanks are extended to Kenneth Matthews of ISMT, Andreas Kerber and Eduard Cartier of IMEC, and Sufi Zafar of IBM for their contributions to this collaboration. ■

References

- [1] P. Packan, *Science* 285,2079 (1999).
- [2] D. Frank, R. Dennard, E. Nowak, P. Solomon, Y. Taur, and H.-S. Wong, *Proc. IEEE* 89, 259 (2001).
- [3] S.-H. Lo, D. Buchanan, Y. Taur, and W. Wang, *IEEE Electron Device Lett.* 18,209 (1997).
- [4] E. Gusev, E. Cartier, D. Buchanan, M. Gribelyuk, M. Copel, H. Okorn-Schmidt, and C. D'Emic, *Proceedings of the Conference on Insulating Films on Semiconductors*, 2001.
- [5] S. Zafar, et. al., *Applied Physics Letters*, 81, 2608 (2002).
- [6] P. Masson, et al., "On the Tunneling Component of Charge Pumping Current in Ultrathin Gate Oxide MOSFETs," *IEEE Elect. Dev. Lett.*, Vol. 20, No. 2, pp. 92–94, 1999.
- [7] S.S. Chung, et al., "A Novel and Direct Determination of the Interface Traps in Sub-100nm CMOS Devices with Direct Tunneling Regime (12–16Å) Gate Oxide," *2002 VLSI Tech. Digest of Tech. Papers*.
- [8] G. Groeseneken, H.E. Maes, N. Beltran, and R.F. De Keersmaecker, "A Reliable Approach to Charge-Pumping Measurements in MOS Transistors," *IEEE Trans. Electron. Dev.*, Vol. ED-31, pp. 42–53, 1984.
- [9] A. Kerber, E. Cartier, et al., "Origin of the Threshold Voltage Instability in $\text{SiO}_2/\text{HfO}_2$ Dual Layer Gate Dielectrics," to be published in *IEEE Electron Device Lett.*
- [10] Y. Kim, A. Agarwal, R. Bergmann, et al., "Conventional n-channel MOSFET devices with polysilicon gate electrode using single layer HfO_2 and ZrO_2 as high κ gate dielectrics," *Technical Digest of the International Electron Device Meeting*, December 2–5, 2001, Washington, D.C., pp. 20.2.1–4.

How to get accurate trap density measurements using charge pumping

The two most common methods used to characterize interface trap state densities in MOSFET devices are charge pumping (CP) and simultaneous C-V (the combination of high frequency and quasistatic C-V) measurement, which is typically done on MOS capacitors. As the size of the transistor scales down, thinner gate oxide is used to maintain proper gate control of the channel. This results in higher gate leakage current due to quantum tunneling of carriers through the thin gate. The higher gate leakage makes characterization of interface traps more and more difficult. Quasistatic C-V becomes impractical for oxide thicknesses less than 3–4nm. Even high frequency C-V measurement becomes a great challenge for oxides thinner than 2nm. The CP technique has much more tolerance than the quasistatic C-V technique, so it can be used for gate oxides thinner than 2nm, using special techniques to correct for excessive gate leakage [1].

After 40+ years of development, interface trap density in silicon dioxide gates is much less a concern than it was years ago, but SiO₂ as a gate dielectric material is approaching its physical and electrical limits [2]. The principal limitation is high leakage current due to quantum mechanical tunneling of carriers through the thin gate oxide [3]. Recently great attention has been paid to the use of high dielectric constant (high ϵ) materials, such as hafnium oxide (HfO₂), zirconium oxide (ZrO₂), alumina (Al₂O₃) and their silicates [4], as replacements for SiO₂ as gate dielectrics. Due to the high dielectric constants of these materials, high gates can be made much thicker than SiO₂ while maintaining the same gate capacitance. The result is lower leakage current—sometimes several orders of magnitude lower. Usually a very thin silicate layer forms between high film and Si substrate. The effect of this silicate layer on interface properties, as well as the charge trapping phenomena inside the high gate, is still to be understood. The CP technique becomes especially useful for characterizing interface and charge trapping phenomena, because the simultaneous C-V technique is very difficult (primarily due to the incapability of the quasistatic C-V technique at the high leakage level) for interface trap characterization.

This article will explain the basic CP technique and its variations and will describe some applications in characterizing charge trapping in high films.

Test procedures and variations

The basic charge-pumping technique involves measuring the substrate current while applying voltage pulses of fixed amplitude, rise time, fall time, and frequency to the gate of the transistor, with the source, drain, and body tied to ground. **Figure 1** shows the

Figure 1. Schematic for charge pumping measurement; source and drain of the transistor are connected to ground; gate is pulsed with fixed frequency and amplitude while body current is measured.

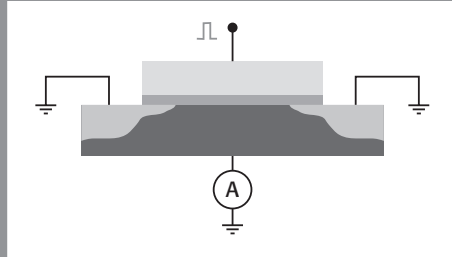
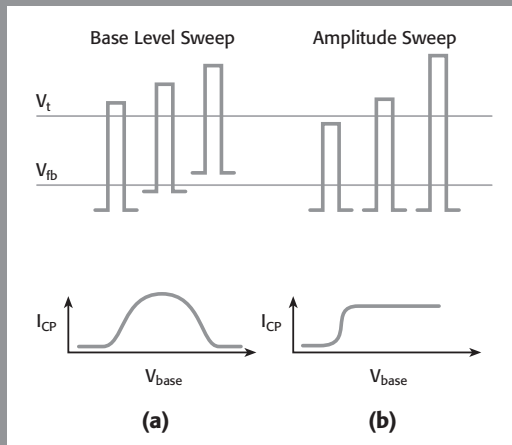


Figure 2. Overview of charge pumping measurements: (a) Pulse waveform for base voltage sweep; pulse amplitude is constant. (b) Pulse waveform for amplitude sweep; base voltage is constant.



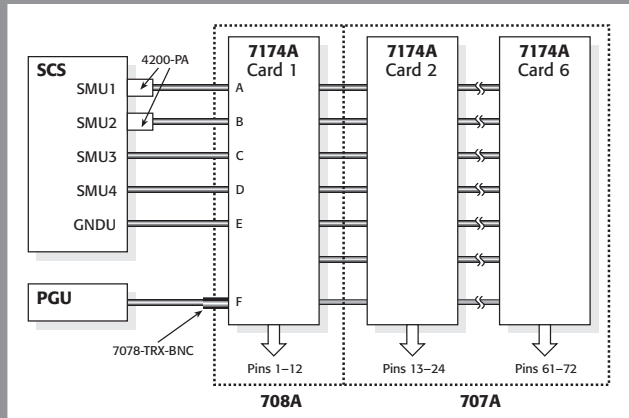
connections. The two most common ways to do charge-pumping measurements are a fixed amplitude, voltage base sweep (**Figure 2a**) or a fixed base, variable amplitude sweep (**Figure 2b**). For high gate stack structures, the CP technique can quantify the trapped charge (N_{it}) as:

$$N_{it} = \frac{I_{cp}}{qfA}$$

(where I_{cp} is the measured charge-pumping current, q is the fundamental electronic charge, f is the frequency, and A is the area) since trapped charge beyond the silicon substrate/interfacial layer can be sensed [5].

In a voltage base sweep, the amplitude and period (width) of the pulse are fixed while sweeping the pulse base voltage. At each base voltage, body current can be meas-

Figure 3.
System setup
for charge
pumping
measurement
with a switch.



ured and plotted against base voltage. The interface trap density (D_{it}) can be extracted as a function of band-bending, based on this equation:

$$D_{it} = \frac{I_{cp}}{qAf\Delta E}$$

where ΔE is the difference between the inversion Fermi level and the accumulation Fermi level [6].

A fixed base, variable amplitude sweep has a fixed base voltage and pulse frequency with step changes in voltage amplitude. The information obtained is similar to that extracted from a voltage base sweep. These measurements can also be performed at different frequencies to obtain a frequency response for the interface traps.

It's relatively easy to perform CP measurements and data analysis using a semiconductor characterization system (SCS) in combination with a pulse generator. **Figure 1** illustrates the connections for a device under test (DUT) with one SMU and a pulse generator without a switch matrix; in **Figure 3** a semiconductor switch matrix is included in the configuration.

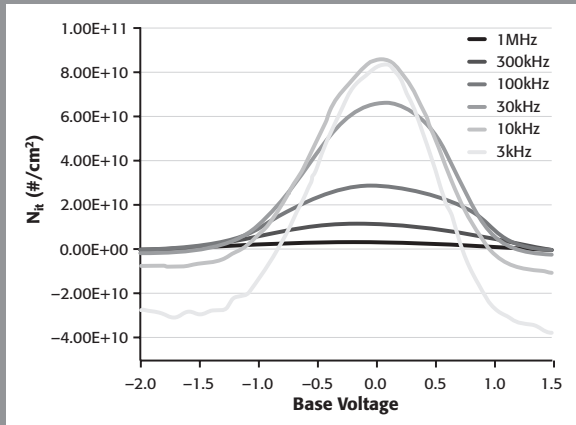
Effects of gate leakage

The measured current (I_m) is the result of averaging of CP current and gate leakage current, as shown in the following. We assume here that gate capacitance is very small, so we can ignore transient currents due to gate capacitance response to pulses. This ceases to be a good assumption if the gate area of transistor under test is relatively large.

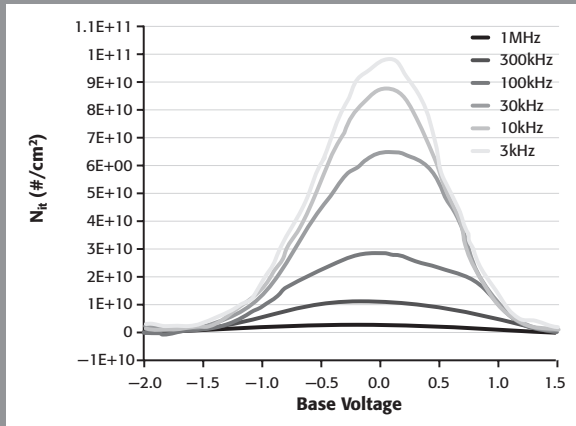
$$I_m = I_{cd} + I_{leak}(peak) \cdot dutycycle + I_{leak}(base) \cdot (1 - dutycycle)$$

Figure 4.

(a) N_{it} extracted without leakage correction, amplitude of gate pulses is fixed at 1.2V.



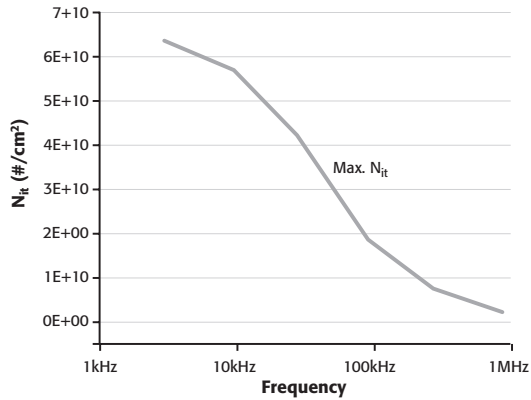
(b) N_{it} extracted with leakage correction.



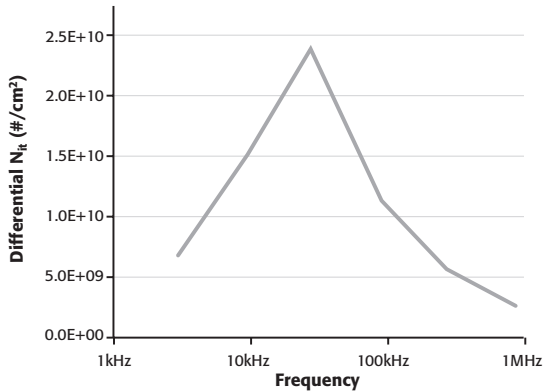
I_{cp} decreases strongly as frequency decreases, but the averaged leakage current is not frequency dependent. By subtracting I_m at very low frequency, leakage current is taken out of the measured I_{cp} . **Figure 4a** shows N_{it} as a function of base voltage sweep at different frequencies. At lower frequency, because the leakage current is of the same order of the I_{cp} , the N_{it} curve is strongly offset. **Figure 4b** shows the corrected N_{it} by using current measured at 300Hz (not shown in the figure). The measurement was done on an n-MOSFET with a base sweep at frequencies from 3kHz to 1MHz with the amplitude of the gate pulse fixed at 1.2V.

Figure 5.

(a) Example of maximum N_{it} as a function of frequency. Because traps responding to high frequency pulses also respond to low frequency pulses, maximum N_{it} at lower frequency is always larger than that at higher frequency. Maximum N_{it} is extracted from Figure 3b.



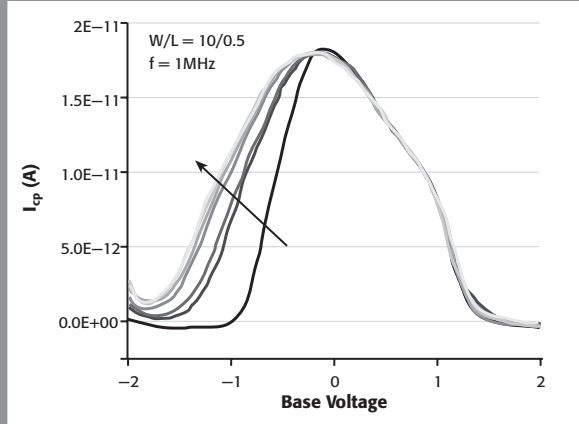
(b) Frequency distribution of measured interface traps. It is extracted by differentiating maximum N_{it} with frequency from Figure 4a.



Frequency dependent trap density

CP measurements at different frequencies yield important information about interface trap density distribution across frequencies. This information is usually shown by plotting either the maximum of N_{it} (cumulative) or the differential of the maximum value of N_{it} as a function of frequency. **Figures 5a** and **5b** show cumulative and differential trap distribution across frequency respectively. Maximum N_{it} was calculated from the N_{it} curve in **Figure 4b** after leakage correction.

Figure 6. Initial trap filling on a “fresh” device measured from CP. The arrow shows increase of I_{cp} as CP is done repeatedly. Amplitude of gate pulses is fixed at 1.2V.



Initial Charge filling

CP can be used to characterize initial stage of gate-channel interface. **Figure 6** shows CP measurement on a “fresh” (never been tested) MOSFET at 1MHz. A series of consecutive CP measurements was repeated. As shown on the graph, the shape of the I_{cp} curve changes as well as the magnitude at lower biases. It then saturates after a number of repeated measurements. This indicates formation of interface traps due to electrical stress imposed by the CP measurement.

Stress-CP measurement

CP measurement is useful alone. It can also be used together with DC or AC stress to study charge trapping as well as new charge creation on the high-Si interface, as well as inside the high film. Charges injected into high film due to stress are calculated by integrating gate current through stress time. $Q_{inj} = \int I_{leakage} dt$. The advantage of stress-CP measurement over traditional stress-C-V measurement (which measures the shift of flat band voltage due to stress) and stress-I-V (which measures the shift of threshold voltage due to stress) is that it can clearly distinguish existing charge trap centers with new centers created by stress [7]. Also, charge relaxation in CP measurement is usually much less than that in stress-C-V and stress-I-V measurements, where trapped charges will relax between stress and the next measurement.

Conclusion

Charge pumping is a powerful tool for characterizing charging phenomena in high gate stacks. Its advantages include requiring less hardware, minimum software interface, relatively simple and easy to perform, good measurement accuracy on interface trap density, and good tolerance on gate leakage. It can characterize initial trap creation (on a fresh device) as well as new traps generated by injected charges (combining stress and CP). ■

References

- [1] S.S. Chung, et al., "A Novel and Direct Determination of the Interface Traps in Sub-100nm CMOS Devices with Direct Tunneling Regime (12~16Å) Gate Oxide," *2002 VLSI Tech. Digest of Tech. Papers*.
- [2] P. Packan, *Science* 285,2079 (1999).
- [3] S.-H. Lo, D. Buchanan, Y. Taur, and W. Wang, *IEEE Electron Device Lett.*, Vol. 18, p 209, 1997.
- [4] E. Gusev, E. Cartier, D. Buchanan, M. Gribelyuk, M. Copel, H. Okorn-Schmidt, and C. D'Emic, *Proceedings of the Conference on Insulating Films on Semiconductors*, 2001.
- [5] A. Kerber, E. Cartier, et al., *IEEE Electron Device Lett.* Vol. 24, pp 87-89, 2003.
- [6] G. Groeseneken, H.E. Maes, N. Beltran, and R.F. De Keersmaecker, *IEEE Trans. Electron Dev.*, Vol. ED-31, pp. 42-53, 1984.
- [7] Y. Zhao, C.D. Young and G.A. Brown, *Semiconductor International*, Oct, 2003.

OVERCOMING THE MEASUREMENT CHALLENGES
OF ADVANCED SEMICONDUCTOR TECHNOLOGIES

SECTION 3

**RF Modeling and
Process Control of
High Performance Analog
and BiCMOS Devices**

RF wafer testing: an acute need, and now practical

RF tests becoming indispensable

Leading semiconductor producers have recently conceded that wafer level RF measurements are acutely needed to develop and produce advanced ICs. To a certain degree, this flies in the face of the 2003 recommendations by the ITRS Technical Working Group for Modeling and Simulation, which states, “The parameter extraction for RF compact models preferably tries to minimize RF measurements. Parameters should be extracted from standard I-V and C-V measurements with supporting simulations, if needed.” The problem is that standard I-V and C-V measurements make the direct extraction of C_{ox} impossible for ultra-thin dielectrics due to high leakage currents and non-linearities. Yet, accurate parameter extraction for HF circuit modeling at 1–40GHz and for RF compact model verification has become essential. This challenge is increasing for high performance/low cost digital, RF, and analog/mixed-signal devices as the industry progresses toward the 65nm node and beyond.

The recommendations for minimizing the use of RF techniques are predicated on the assumption that they cannot be made effectively, particularly in a production environment, which may have been the case in the past. However, new parametric test systems now make fast, accurate, and repeatable RF parameter extraction almost as easy as DC testing. In fact, one system can take precise DC and RF measurements simultaneously, making it suitable for both lab and production use.

RF testing apps

Whether you are manufacturing RFICs on III-V wafers for cell phone modules or high performance analog on silicon-based technology, predicting final product performance and reliability requires wafer level RF s-parameter measurements in development and production. These measurements are an important addition to DC data in forming a complete picture of device characteristics. They also offer significantly more information with fewer measurements than a DC-only test suite.

Power amplifier RFICs are an obvious candidate for high frequency testing. They are extremely complicated, yet susceptible to end market pricing pressures. This makes them highly sensitive to testing costs in production, where functional tests are conducted under low bias conditions from 1GHz up to 40GHz, depending on their design and application. RF measurements have been limited to functional tests of packaged parts at the end of the line, as this testing is perceived as high cost and problematical in terms of repeatable, accurate results.

IC fabricators can also use RF wafer level measurements to extract figure of merit parameters on various high performance analog circuits at the 180nm node and beyond. SOCs that combine memory with RF, analog, and high speed digital devices have comparable RF test requirements.

Characterizing equivalent oxide thickness (EOT) on high D gate dielectrics is critical in high performance logic devices at the 130nm node and beyond, in the development of new materials, and for continued scaling of future IC generations. For example, RF measurements can play an important role in accurate modeling of dielectrics and their behavior in MPU, ASIC, FPGA, and DSP devices. This has been done on prior generations of technology using multi-frequency capacitance measurements and with advancing technology there has been a shift to high frequency capacitance (HFCV) measurements. However, HFCV is inadequate for ultra-thin dielectrics, one reason being that the HFCV instrument (not the DUT) introduces a series resistance into the measurement.

Challenges in standard I-V and C-V measurements

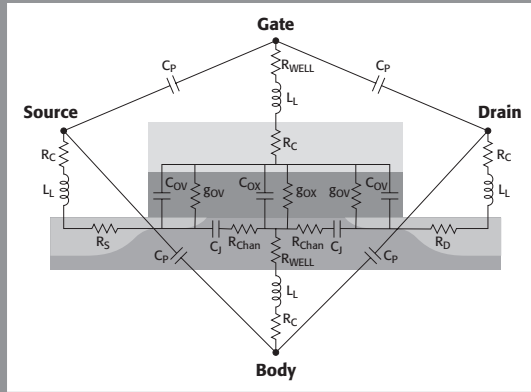
RF parameters extracted from s-parameter data are included in compact simulation models used by design engineers during product development. However, manufacturers have taken wafer level s-parameter data only in device modeling labs, due to the complex nature of the measurements and associated cost. Even in a lab environment, accurate extraction of RF parameters has been a challenge due to the sensitivity of the measurement to probe contact resistance variation. Making the large number of measurements required for production monitoring quickly has seemed impractical, if not impossible.

One of the main technical challenges in RF testing is accurate de-embedding of the DUT and measurement system. This is particularly true in the extraction of C_{ox} when characterizing ultra-thin gate dielectrics. **Figure 1** is a simplified circuit model of a real transistor that shows the components which complicate direct measurement of C_{ox} . These components include overlap capacitance between the gate contact and the source/drain well, gate resistance, lead inductance (from DUT to contact pads), contact resistance (between probe needle and contact pads), and channel resistance. These components must be segregated by the appropriate extraction after de-embedding. Correction algorithms are applied for contact resistance, lead inductance, and probe pad capacitance.

Reluctance in adopting RF C-V

These difficulties have important implications in a production environment. Unreliable measurements can hinder yield management. A bad measurement result on a good device is referred to as alpha error. In a production environment, this may mean that a wafer has been improperly scrapped. The misleading ITRS information and the slow, painstaking process that many companies experience in their modeling labs combine

Figure 1. Simplified circuit model of a MOSFET DUT. The C_{OX} measurement factors to consider are parasitic capacitance between contact pads and leads (C_P), contact resistance (R_C), lead inductance (L_L), channel resistance (R_{Chan}), and overlap capacitance (C_{OV}).



to make engineers reluctant to adopt production RF measurements, believing they will have high alpha error. It is also perceived that throughput and operational costs will be unacceptable and that a high level of technical support is required to interpret results. Low throughput on prior generations of RF systems resulted from calibrations and measurements needing to be repeated due to contact resistance problems. Calibrations on these older systems also did not hold for different measurement frequency sets. High operational costs are associated with manual probing of gold calibration standards, which have soft pads and expensive RF probes that wear out quickly with over-scrubbing. There is also the false perception in the market that a special prober or chuck is required for wafer level s-parameter measurements. These factors result in a high perceived cost of ownership for RF C-V and reluctance by users to adopt the solution.

Succinctly put, the industry's overall perspective on wafer level RF testing has been that it is complicated and expensive. This is based on a view of RF measurements as black art and that implementing them in a high volume fab, run by a production floor operator, is fantasy. Additional concerns regarding RF measurements in production are that:

- Extensive test structure changes are needed.
- Results are unstable, varying tool-to-tool, operator-to-operator, and day-to-day.
- RF specialists must baby-sit every tool.
- Substantially different lot routing and operational workflow may be required.

- It is doubtful this can be a real-time technique.
- Lab grade results are unlikely.

Nevertheless, by maintaining the status quo based on these perceptions, fabs are “flying blind” in the implementation of new designs and processes for RFICs, new gate materials, and other advanced devices. The consequences are design and process iterations that greatly increase costs and time to market, accompanied by lower initial yields.

Third Generation Parametric Testers Provide a Solution

The key to making wafer level RF testing a production process control tool is fully automated measurements. This means that a robot delivers the wafer, the calibration standard, and the probe card to where they are needed. In other words, a major test system design goal is absolute data integrity without human intervention. If intervention is required, it should be accomplished by the fab host or the test controller, based on intermediate test results or operational requirements.

Third generation testers now available have features that allow this type of operation to 40GHz. Being designed specifically for a production environment, they avoid some of the attributes of testers designed for lab use. Lab instrument design focuses on optimizing the manual use-case and features other than those associated with production. However, users of these instruments pay for every extra gigahertz and other incremental features. There is no upgrade path to support changing needs from 6GHz to 65GHz as applications change. Third generation testers support this upgrade path.

Third generation testers address the need to automatically de-embed and extract the measurements according to the DUT characteristics, which is a major technical challenge in getting reliable C_{ox} results. These algorithms, coupled with improved interconnect technology and automated calibration procedures, allow fast and accurate RF parameter extraction from s-parameter measurements.

Correcting random measurement artifacts is a prerequisite to accurate de-embedding. For example, any change in contact resistance in a system with 50 Ω characteristic impedance limits repeatability. Instrumentation manufacturers must identify all the sources of instability in RF measurements and design the test system to avoid them. Innovative design of the system interconnections is required to provide repeatable links between major system components.

Automatically measuring probe contact resistance and adjusting probe overdrive is another way an instrument manufacturer can assure repeatable measurements. By measuring the actual value of the contact resistance before RF measurements are taken, they can be corrected for the value of contact resistance, especially important for pas-

sive devices. Another benefit of actually measuring contact resistance is the ability to automatically initiate probe tip cleaning when the resistance gets too high due to contamination. Good overdrive control and cleaning only when needed will increase probe life significantly, which reduces a major consumable cost. (RF probes cost about \$1000 each.) This should also be part of the statistical process control of the tester.

With stable and known parasitics, the Smith chart curves generated from collected data are free of artifacts; there is no need for specialists to analyze and interpret results. In older systems, an expert in RF measurements was required to monitor data (i.e. curve traces of every measurement set), look for strange or unexpected results, and then analyze those results to make sure they represented process variations, instead of measurement anomalies.

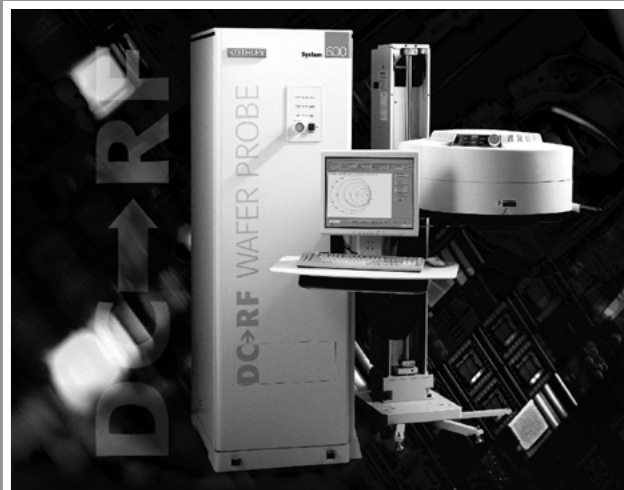
Improved logic in third generation parametric testers makes continuous monitoring of RF measurement quality a reality and reduces or eliminates the need for support by RF specialists. With these systems, different production floor operators can get repeatable real-time results across a wide range of products and production tools. RF measurements are almost as easy as making DC measurements, which are also required to completely characterize wafer devices. In fact, one third generation system can make DC and RF measurements simultaneously. (See sidebar.) This system contains a number of other refinements that speed up throughput, making it practical to do high volume wafer level testing for process monitoring and control. These same features speed up measurements in the modeling lab without sacrificing lab grade results, thereby shortening the development cycle and time to market. All this can be done without purchasing special probes, through easy system upgrades. When the calibration standard is stored on the prober, the operational work flow is identical to DC-only testing and is changed only during periodic maintenance cycles.

Innovative designs for RF testing

For many years, RF parametric testing at the wafer level was the province of “big iron” ATE systems limited to 6GHz or less or lab systems that were not capable for production use. Both were impractical for statistical process characterization and monitoring. To solve these and other problems associated with existing systems, in 2001, Keithley introduced its DC/RF series of parametric testers.

With these systems DC and RF testing can be done in parallel, asynchronously. This means that DC tests can be run in the background as RF testing is performed or visa versa, depending on which type of testing has the more complex attributes. As soon as either set of measurements is complete, the system is ready to perform more tests. Empirical data shows that neither DC nor RF test results are affected by running all test types in parallel. Since DC and RF measurements can be made simultaneously with a single prober insertion, throughput is greatly increased. System software provides real-

Figure 2.



time de-embedding and parameter extraction, while its mature point-and-click GUIs are unmatched by other systems in their ease of use.

The Keithley system design was influenced by a successful collaboration with several customers, resulting in innovations in interconnect and VNA integration, as well as adapting the parametric tester invented by Keithley to accommodate RF measurement needs. Since then there has been continuous improvement in all aspects of these systems, and in July 2003, Keithley received the Attendees Choice award at SEMICON-West for the 300mm production capability that was added.

Recently, Keithley introduced the third generation of these DC/RF parametric testers (**Figure 2**) and already has several application successes with prominent semiconductor fabs. In these applications it was demonstrated that different technicians using the same system in a production facility could get the same results; in the past, this hasn't even been possible in the lab on other vendor's tools. When used in the lab, the Keithley tools provide excellent correlation with measurements made in production, even though the lab test suites are far more complex.

From the outset, a design objective was to make DC/RF parametric test systems that were usable by fab equipment operators to get high quality RF measurement results. Now it's possible for ordinary operators to push a wafer boat into the test system and get lab-grade results without specialized RF training. All the operator has to know is the name of a parameter; they do not need to be concerned about calibration, de-em-

bedding, or parameter extraction techniques. This can all be done at high production throughput, not at rates normally associated with lab measurements.

These capabilities are the result of patented and patent pending calibration procedures, interconnect technology, and de-embedding algorithms. Included with each system is the industry's largest RF parametric extraction library.

The Keithley systems are the only production solutions available for RF C-V measurements to meet the needs of the 65nm node and beyond, and those solutions have been qualified for 300mm production environments. At this writing, they are qualified for RF production test at seven companies in the US, Europe, and Asia. These applications include on-wafer RF functional testing of IC devices at two different manufacturers of cell phones chips. Keithley is the only RF test system supplier qualified for production measurements higher than 6GHz. Four different probe cards have been qualified for these systems.

While these systems were designed primarily for a process integration interface, where modeling meets parametric test in the fab, their high precision, data integrity, and throughput makes RF measurements practical for modeling labs, parametric production monitoring, and end-of-line functional testing. For instance, one modeling lab using the Keithley system reported that data collection and analysis that formerly took up to 13 weeks could be completed in as little as one eight-hour work shift. The large quantity of high quality data being collected effectively closes the "model-to-measurement" gap that has previously existed. It is now possible to verify new RF process models in less than a week, compared to more than two months using older RF solutions.

High speed, high quality data collection is largely the result of self-monitoring features in these systems. For example, they monitor for human triggered events, such as an undocked test head, any change in equipment due to movement or reconnections, and calibration not being initiated at the proper interval (usually, a three-day span). In all cases, the system recalibrates itself automatically, which only takes about two minutes.

A SofTouch (automatic Z adjustment) control feature results in superior RF measurements and lower consumables cost. In the past, the technique has been to overdrive the probe to scrub through contact surface resistance, but you still didn't know the magnitude of that resistance. The Keithley systems measure the contact resistance and limits the amount of overdrive and probe wear. Additionally, the value of contact resistance is used to correct the measurements.

The result of overdriving is shorter probe life; in other systems, the best achievable RF probe life is about 3000 touchdowns. This is drastically improved with SofTouch control. In one application the customer is getting a useful life of up to 300,000 touch-

downs; in another, the user is getting up to four million touchdowns on a set of RF probes. In the latter case, the savings from fewer probe replacements over a six-month period repaid the cost of an RF upgrade to a DC only system.

Another aspect of Keithley's probe control is better utilization of prober overhead time, resulting in higher throughput. While the prober is indexing and the needles are in the air, the system makes s-parameter measurements to determine if needle tips are getting contaminated. If so, the probes are moved to the cleaning pads for cleaning. The ability to automatically trigger probe tip cleaning and calibration as needed, within a single test execution thread or single command from a 300mm host, is unique to Keithley's third generation systems. Moreover, this function is executed without requiring a unique configuration in the test program, without any delay or disruptive communication with the 300mm host, and without the robot having to do anything. The system also does data extraction while the needles are in the air. The only thing it does when the probes are down is make measurements. Thus, no CPU cycles are wasted.

Probe card change-outs have long been a problem in many parametric test systems, and the cards must be changed every time there is a new type of wafer (different product) coming through the fab line. These problems are exacerbated in the case of RF testers, where mechanical damage during change-out is a frequent occurrence. For example, in many of these systems mechanical interconnections require the use of a cumbersome torque wrench. This frequently results in accidental damage as the wrench is dropped on probes, or due to over-torque damage, or inaccurate calibration occurs due to under-torqued connections. Even without these problems, a technician's hands can come in contact with the probes and bend them or damage other probe card parts.

These problems are avoided in the Keithley system: an operator simply pushes a button, the probe card comes out and is removed, and a new one is dropped into the slot, all of which can actually be accomplished by a robot instead of the operator. All the while, the test head stays docked, so calibration is often not affected for most probe types. Automatic probe card change for 40GHz measurements is another unique feature of the Keithley systems.

Another common problem in older wafer level RF testers has been oversized bias tees. These tees are used as part of a Kelvin connection to supply DC bias to the probe. Some of these tees are as big as a fist, making them hard to fit them into a probe head, and prone to interconnection problems that lead to measurement error. Working with Anritsu, Keithley developed a miniature Kelvin bias tee that can be located in the test head for stable connections with minimum parasitics.

With the flexibility to upgrade Keithley legacy systems going back almost 20 years and use most production probers up to 40GHz, Keithley's DC/RF system design pro-

vides the lowest cost of ownership available. Even a new system provides an attractive cost of ownership because of its initial price, high throughput, low operational/support costs, and low consumables cost.

Beyond the hardware and software in its DC/RF parametric testers, Keithley engineers work closely with customers to apply these innovative systems and RF measurement techniques to particular applications. Keithley develops extraction libraries that accommodate particular device and fab operation subtleties, helping users really understand what RF measurements mean in terms of semiconductor device processes. Keithley systems are also being used in labs to create statistical models much faster by collecting an extraordinary amount of reliable data in a much shorter time than ever thought possible. All this is being done with DC femtoamp precision, and RF measurements up to 40GHz. ■

Statistical process control of wireless device manufacturing requires production worthy s-parameter measurements

RF process monitoring – who needs it?

Almost without exception, semiconductor device manufacturers use DC data for statistical process control of their manufacturing operations. They have traditionally taken s-parameter data only in device modeling laboratories due to the complex nature of the measurements and associated cost. RF parameters are extracted from the s-parameter data and included in the simulation models used by design engineers during product development. RF parameters for modeling and DC data for production was a working paradigm until product performance approached the gigahertz range. Process control for gigahertz devices requires RF parameter sampling to meet the Known Good Die goals of RFICs for cell phone modules, guarantee the frequency performance of DSP chip level interconnect, and monitor gate dielectrics with complex material properties. As device performance improves to meet the demands of the wireless market, component manufacturers are struggling to improve test coverage without unacceptable increases in test cost. Several leading manufacturers have attempted to migrate the RF parametric measurement capability from their device modeling labs to manufacturing operations with little success and high levels of frustration.

Traditional methods of RF parameter testing

As might be expected, RF testing began in product development to describe device performance in terms of familiar transmission line characteristics, such as 2-port/4-terminal s-parameters (s_{11} , s_{12} , s_{21} , and s_{22}) and noise parameters (NF, 1/F Noise). The instruments used for microwave device testing were added to equipment racks containing DC test equipment to allow measurements from DC to RF frequencies. However, these rack-and-stack test systems have traditionally required multiple probe insertions on the wafer for both calibration and measurements. A large number of individual measurements are required to characterize a device fully at DC and RF frequencies, so this type of system design results in extremely long test times. For example, with a Vector Network Analyzer, it can take from several minutes to a few hours just to complete the initial calibration of the system. Measurements are rarely made after only a single calibration with these types of setups.

The long calibration and test times associated with traditional rack-and-stack systems are a result of manual methods that require substantial operator intervention to

verify test system performance. Although the RF test methodology is well established, actual implementation is complicated and may require a practitioner with Ph.D. credentials to obtain accurate calibrations and measurement results. It can take several days to complete a full battery of tests using traditional equipment and methods. This may be tolerable during product development for initial device characterization, but is not practical in a process monitoring environment. (See **Table 1.**)

Table 1. Characteristics of a traditional rack-and-stack DC/RF parametric test system.

Special probe and chuck required.
No test executive; each test program is usually created in basic.
Multiple manual calibrations needed; low repeatability.
Manual de-embedding ¹ prone to probe variations due to contamination (resonance).
Slow measurements, data transfers, and extractions.
Expert user required to extract RF parameters from s-parameter data.
Low data integrity leads to repeat measurements and low productivity.
High cost of consumables (e.g., probe tips).
Not practical for production monitoring, only device characterization.

¹ De-embedding refers to a calibration procedure that removes the effects of parasitic impedance associated with probe pads and interconnections. These effects can include the generation of resonant frequencies that obscure RF measurements at certain frequencies.

Fortunately, there are no fundamental impediments to automating RF parametric test algorithms. With appropriate instruments and probe hardware, plus test executive software, calibration and test times can be shortened dramatically. By pushing individual measurement times down to the millisecond domain, RF parameter extraction becomes practical for process monitoring.

Automated probe package hardware and software

To ensure high reliability and repeatability, the probes, probe cards, and interconnect used with the automatic probe station must exhibit low loss and reflections from DC to the highest frequency of interest. With today's wireless devices, that upper limit may be 40GHz. For such broadband semiconductor testing, the interconnection scheme typically uses precision miniature 50 Ω coaxial cable from the probe tips to the connector interface. This coaxial design provides lower loss and less radiation than coplanar designs, plus measurement repeatability with about -80dB isolation.

Repeatability also is a function of stable calibration, which is related to VNA architecture, probe design and probe package software. Independently spring-loaded beryllium-copper or tungsten probe tips assure repeatable contacts at the probe points, minimize circuit damage, and increase probe life. The best designs allow 300,000 or more touchdowns before repair is required. The probe control algorithm used with

this hardware should automatically provide a small amount of overdrive so the probe point does not over-scrub the connection pad surface, but makes a reliable contact that is free of dust, dirt, and oxide contamination. The slight scrubbing and ability to view the exact contact area eases probe positioning and allows the precision necessary for good calibrations. For enhanced data integrity, some probe card algorithms include a burnishing routine that supports automated cleaning of probe tips.

With an automated calibration algorithm in the test executive, a system level calibration can be as short as two minutes. By calibrating the instruments, interconnects, probe card, probe card adapter, and calibration substrate as a complete system, the highest overall accuracy is assured. This also eliminates the need to recalibrate for changes in test frequency or the number of data points, which consumes a lot of time with manual calibrations.

Use of a VNA for RF measurements

The heart of the s-parameter measurement system is the Vector Network Analyzer (VNA). The primary considerations for this VNA are noise floor ($<-90\text{dBm}$), how the sweep is synthesized (lock on every frequency point), high speed data transfer, internal mass storage, software initiated calibration, and high MTBF ($>50,000$ hours). A VNA can provide a high level of precision in auto-reversing s-parameter measurements of active and passive multi-port devices at microwave frequencies. Within 350ms, the VNA can collect 100 or more data sets with 1kHz resolution. If the test executive contains an appropriate library of RF macros, the VNA s-parameters can be used to extract corrected RF parameters quickly (ft, fmax, fmin, Q, C, Rb, L, Load Pull, and many other RF characteristics). These can be presented in tabular form, as smith charts, X-Y plots, etc. See **Figures 1 and 2**.

Keithley Sx00DC/RF Series

Just as Keithley Instruments was the first to bring low current DC measurements to the production test environment, we have used the design approach described in this paper in the new Keithley Sx00DC/RF APT systems for process monitoring. In partnership with Anritsu Corporation and GGB Industries, Keithley developed a low cost, merged DC/RF test solution based on the best technologies available. Key performance features include:

- Keithley Sx00 APT based DC test system with solid state transfer switch and Anritsu VNA technology for fast RF measurements with 1kHz frequency resolution, stable auto-calibration, and a modular upgrade path from 20–110GHz.
- Anritsu VNA includes integrated fast sweeping source, auto-reversing s-parameter test set, and four-channel receiver.

Figure 1.

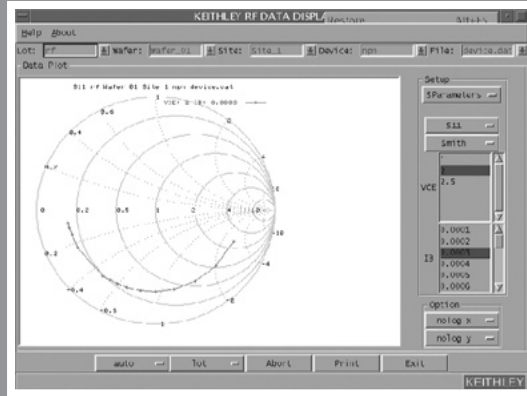
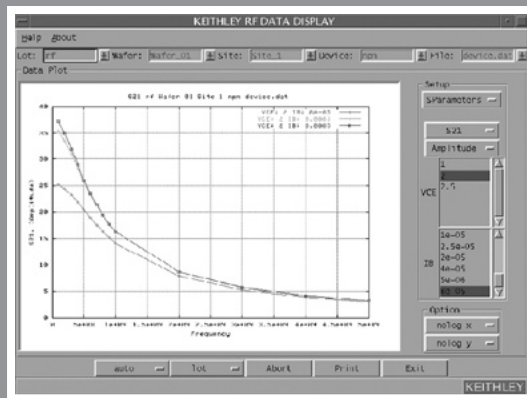


Figure 2.



- GGB self-leveling G-S-G probe technology provides measurements from DC to 200GHz. With 300,000 or more touchdowns (a 100× increase over older systems).
- Keithley System 41 RF Switch technology is an option for multiple DUTs for measurements up to 20GHz.
- Both single and segmented limits can be used for PASS/FAIL testing; 1 to 4096 averages can be performed on the data.

- Full factory automation, efficient test programming, and high productivity tools provide high measurement throughput and accuracy for all DC and RF parameters.

This APT system design can be used with any automatic prober, and provides fully automated single-pass calibration that is quickly executed during testing—without the need for human verification. The calibration includes automatic de-embedding of probe pad/interconnect impedance that would impair data integrity, and can be completed in approximately two minutes. (The manual calibration procedures of traditional rack systems take up to four passes and two hours. Furthermore, manual de-embedding is prone to probe variations due to contamination, resulting in low data integrity that leads to repeat measurements and reduced productivity.) Automation of probe tip cleaning is also accomplished with this system.

Sx00DC/RF test functions are integrated with the Keithley test environment (kte) production productivity tools. With the kte user library of RF macros, even novice users can quickly extract RF parameters. The result is 3× better raw measurement speed with real-time data extraction and binary transfers to file or a plant network. A data-driven programming environment and kte's user access points allow easy program adaptation, even during test. With the fast RF auto-calibration (which is unique to the Keithley system), overall system throughput can be increased by 10× compared to traditional rack system designs.

This high speed and repeatability makes the Sx00DC/RF's use practical in process monitoring. However, single insertion DC/RF testing requires compatible test structures. Therefore, Keithley offers test structure design support (test programs included) to facilitate this new capability. ■

OVERCOMING THE MEASUREMENT CHALLENGES
OF ADVANCED SEMICONDUCTOR TECHNOLOGIES

SECTION IV

Reliability Testing

Wafer level reliability testing— a critical device and process development step

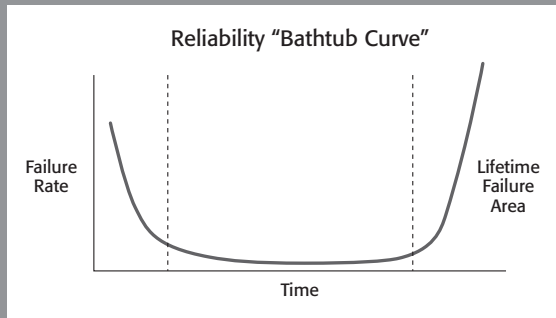
Changing reliability issues

The continuing push for more devices on a chip and faster clock speeds is driving the demand for shrinking geometries, new materials, and novel technologies. All these have a tremendous impact on the lifetime and reliability of individual devices due to increased fragility, higher power density, more complex devices, and new failure mechanisms. Processes that once produced devices with 100-year lifetimes may now yield only 10-year lifetimes—uncomfortably close to the expected operating life of systems using these devices. The smaller margin of error means that lifetime reliability must be designed in from the start and constantly monitored, from device development, through process integration, and into production; even small lifetime changes can be catastrophic to today's devices.

While reliability testing can be done at the packaged device level, many IC makers are migrating to wafer level testing for a number of reasons. Wafer level reliability (WLR) testing allows testing earlier in the process, eliminating much of the time, production capacity, money, and material lost if the packaged device fails. Turn around time is much less as a wafer can be pulled directly off the line and tested without the delay of sending the part away for packaging, which can be up to a two-week process. Much of the testing is the same, allowing for relatively easy migration to wafer level testing.

Stress-measure testing is a common technique used to evaluate operating lifetimes and wear-out failure mechanisms in semiconductor devices. This testing is focused on

Figure 1. Typical semiconductor reliability curve.



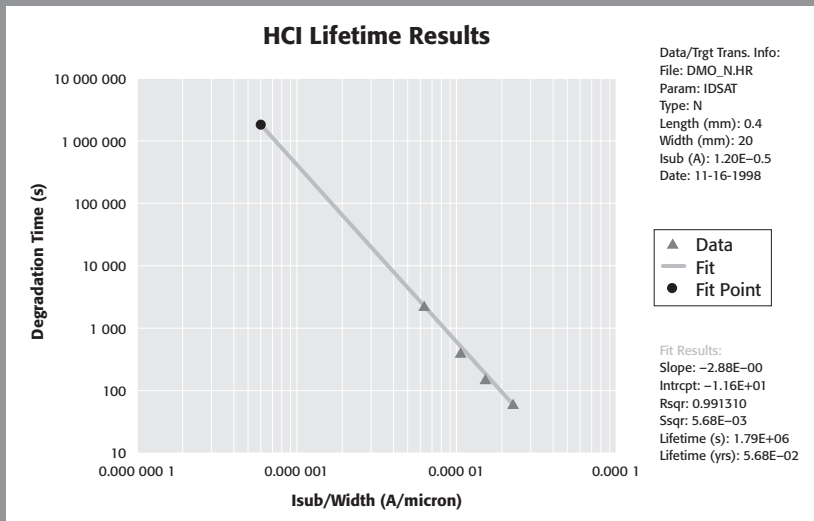


Figure 2. Example of lifetime reliability extrapolation from HCI testing.

failures on the right side of the typical failure rate bathtub curve (**Figure 1**) – i.e., failures not associated with infant mortality or manufacturing failures.

A stress-measure test is used to generate curves quickly that can be extrapolated to predict the expected operating lifetime of a device. This data is used to evaluate the device design and monitor manufacturing processes. Since typical device lifetimes are measured in years, techniques are needed to accelerate the testing. The most efficient method is to over-stress the device, measure degradation trends of key operating parameters, and extrapolate the data to the full lifetime. For example, the lower right portion (collected data) of the curve in **Figure 2** was generated using high stress conditions. The data generates a line that can be used to predict device lifetime under normal operating conditions (upper left portion of the curve).

Common WLR tests that use stress-measure techniques have included Hot Carrier Injection (HCI) or Channel Hot Carrier, Negative Bias Temperature Instability (NBTI), Electromigration, and Time Dependent Dielectric Breakdown (TDDb) or Charge to Breakdown (Q_{BD}). These tests have become critical in mainstream CMOS device development and process control. However, new scale factors and materials now require modifications to these established techniques, and demand instrumentation features that can implement these new techniques.

Hot Carrier Injection

HCI has been one of the key reliability tests in the last couple of CMOS generations. This is a process where high lateral electrical fields in a MOSFET generate hot carriers (high energy electrons or holes) that can damage the MOS gate oxide interface and degrade the device's I-V characteristics. This phenomenon gets worse as channel length decreases, because the lateral electric field in the channel is a function of gate voltage divided by channel length. As channel lengths have been decreasing proportionally faster than gate voltage, the increases in lateral electrical fields are causing higher energy carriers and more potential damage to the gate oxide. This damage is due to the high kinetic energy of accelerated carriers that produce electron/hole pairs through impact ionization.

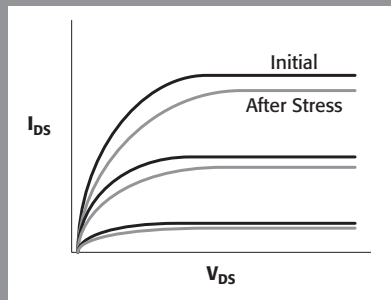
Degradation will be seen in the device's I_{DS} (**Figure 3**), transconductance, and threshold voltage. Degradation first slows down the operation of the device, and will eventually cause it to stop working all together. The HCI test measures how fast a MOSFET transistor degrades when voltage stress is applied, and uses stress conditions to accelerate the degradation for quicker results that can be extrapolated to lifetime predictions under normal operating conditions (**Figure 2**).

Negative Temperature Bias Instability

NBTI is a failure mode that is problematic in PMOS transistors and getting worse as threshold voltage continues to drop. NBTI degradation is measured by time dependent shifts in threshold voltages, and is associated with slower operation, more leakage, and lower drive current under negative bias stress at high temperature.

The NBTI test is typically a stress-measure sequence loop. During the stress, negative gate bias voltage is applied with the rest of transistor terminals grounded. Between two consecutive stresses, drain current is measured at normal operating condition. Degradation of drain current or threshold voltage is plotted as a function of stress time.

Figure 3. I-V curve showing HCI induced I_{DS} degradation after voltage stress.



All the stress voltages and subsequent measurements are done at high temperature (for example, 135°C).

A unique characteristic of NBTI degradation is that it can relax when stress is off. When gate voltage stress is turned off, the degradation of drain current and threshold voltage may recover and change back toward their original value. The rate of recovery is strongly dependent on temperature. At room temperature, as much as 100% recovery has been reported. If stress is resumed on the gate after recovery, the degradation will follow the previous degradation curve. At higher temperatures, there will be a portion of the degradation that is irreversible. This is called degradation lock-in.

Another aspect of the NBTI recovery problem is associated with typical transistor operation, where it is turned on and off very often. When the transistor is off, NBTI degradation may recover. Therefore, if one uses the traditional DC stress and degradation technique, there will be no recovery effect and it may underestimate the lifetime of the transistor.

One approach to handling this recovery dynamics problem is to use pulse stress instead of DC stress. In this technique, the transistor is biased alternately between stress and a normal operating condition. Then degradation of threshold voltage is measured as a function of pulse frequency. This test routine provides some very important information about the nature of recovery in different applications. For example, the switching frequency is not the same for transistors in different circuits with different functionality. The frequency dependency of NBTI degradation may reveal that some part of a circuit will fail before the rest.

New Reliability Techniques

Device reliability is increasingly tied to design issues and process damage that involve the silicon (Si)/gate dielectric interface. The gate dielectric is the most sensitive part in a MOSFET. Charges inside the gate oxide and at the oxide/Si interface greatly affect transistor performance, such as when and how fast transistors turn on. Threshold voltage is directly related to the amount of charge inside the gate, and between the Si and gate interface. Damage to the oxide and interface during circuit operation is the main source of reliability problems. Channel hot carrier induced degradation, negative bias temperature instability, and charge trapping all come from the result of interface and oxide damage. Characterization of interface/oxide degradation is key to understanding those device reliability issues.

Charge Trapping in High κ Gate Dielectrics

While high κ material can help solve ultra-thin gate leakage problems with leading edge processes, there is no free lunch. Associated with this advantage are several technical hurdles that must be overcome. One is that the quality of the film is not very

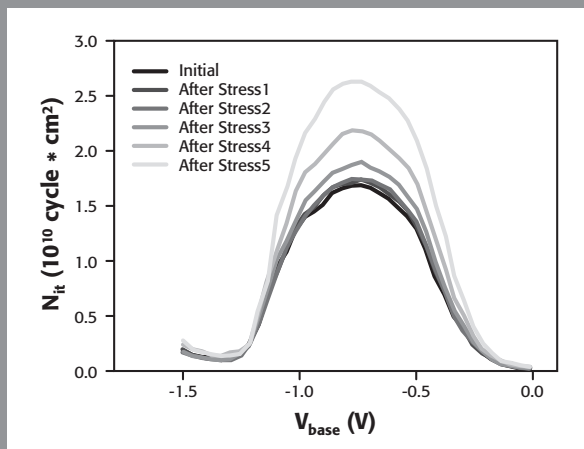
good. There are a large number of interface states as well as traps in the bulk film. The problem is, when the transistor is turned on and carriers flow through the channel, some of these carriers will be trapped in the interface and bulk of the film, resulting in a shift of the threshold voltage. This charge trapping problem is reportedly more severe for NMOS than for PMOS, since electron trapping is much easier than hole trapping.

Besides looking at charge trapping during normal transistor operation, one can stress the gate so that charges are intentionally injected into it (charge pumping). The purpose of doing this is twofold: (1) it's possible to control the amount of injected charge; (2) it allows one to see if there is any interface damage due to stress and how the damage affects charge trapping behavior. The damage to the interface can be seen when measuring charge pumping current after each stress. **Figure 4** shows that accumulating stress creates more interface states.

Reliability Test Instrumentation Trends

As the previous text indicates, reliability tests have evolved to match the needs of new device designs and materials. While HCI is still an important reliability concern, engineers now worry about NBTI for PMOS, charge trapping for high κ gate transistors, and cross effect between NBTI, TDDB and HCI, such as NBTI enhanced hot carrier, and TDDB enhanced NBTI. To deal with these new phenomena, measurement methodology has evolved from DC stress and measurement to a point that both DC and pulse stress are used to study degradation relaxation effect. Furthermore, instrumentation now includes more comprehensive device characterization suites, which include DC I-V, C-V, charge pumping, and charge trapping.

Figure 4. Formation of interface states after DC stress.



These evolving test requirements are challenging engineers to find the right instrumentation for efficient device and process development. The tool selected should be sensitive enough to capture all the pertinent details of parameter degradation due to stress, and flexible enough to adapt non-traditional WLR tests, such as stress C-V, charge pumping, etc. This tool should also be extendable so that one does not need to buy a completely new system every time a new test issue comes up. Finally, the tool should be easy to use so that one can focus valuable time on interpreting data, not learning to use the test system.

In terms of features, a modern reliability test stand must provide the following:

- Hardware and software that accelerates testing without compromising accuracy and extrapolated lifetimes.
- Semi-auto or auto-prober with a thermal chuck.
- Manipulators or a parallel probe card with low leakage.
- Drivers to control instruments, probers, chucks, create tests, execute tests, and manage data.
- Flexibility to accommodate user-changeable tests and stress sequences for new materials and failure mechanisms.
- Analysis software that provides easy extraction of final lifetime predictions from accelerated short-term tests.

Keithley Solutions

To meet these emerging requirements, Keithley introduced the Model 4200-SCS Semiconductor Characterization System with reliability test enhancements. Using the Model 4200-SCS, engineers can easily put different measurement techniques together to collect data in a timely fashion. This system is configurable from two to eight SMUs. The optional preamp has 0.1fA resolution. The Model 4200-SCS can also control other instruments, such as a switch matrix, C-V meter, and pulse generator without user programming. This can be done using GPIB, Ethernet, or RS-232. The interactive software has a test plan manager, interactive test setup interface, Excel-like data sheet, easy graphing capability, and more. The Model 4200's flexibility makes it an ideal characterization tool, whether it is used for development in an interactive manual mode (for single test operation) or in more automated use cases.

The Keithley Model 4200-SCS with KTEI software makes reliability testing quick and easy. Its intuitive point-and-click interface, combined with built-in test routines, makes setting up and running reliability tests as easy as getting I-V curves. The Model 4200-SCS comes with a variety of standard stress-measure tests ready to run on bootstrap, which are

easily modifiable for customized testing. Its industry leading sensitivity and low-level measurement capabilities make it ideal for accurately tracking the smallest change in a degrading parameter. One feature that turns the Model 4200-SCS into an ideal reliability development station is the enhanced wafer level stress-measure loop built into the test plan manager. It includes a stress-measure loop with exit logics and a site loop for stepping through sites on a wafer. By taking advantage of these two loops, one can easily set up a customized wafer level reliability test without any programming.

Figure 5 shows the bundled HCI sample project. On the left is the sequencer showing the order of measurement tests and the overall structure of the project. The gray highlight is on the stress portion of the sequence, where stress values, time values, and looping control are set up. Listed below the stress piece are individual tests for monitoring specific parameters. The graph shows a particular parameter being tracked over time, with each dot representing a different measure cycle after a stress cycle.

In addition, when performing traditional WLR tests such as HCI, electromigration, and TDDDB/ Q_{BD} , the biggest advantage of the Model 4200-SCS is that it can easily include customized WLR test routines. For example, the voltage waveform shown in **Figure 6** illustrates different test modules, such as C-V, I-V and charge pumping that can be simply combined in the test plan manager for a looping sequence without programming.

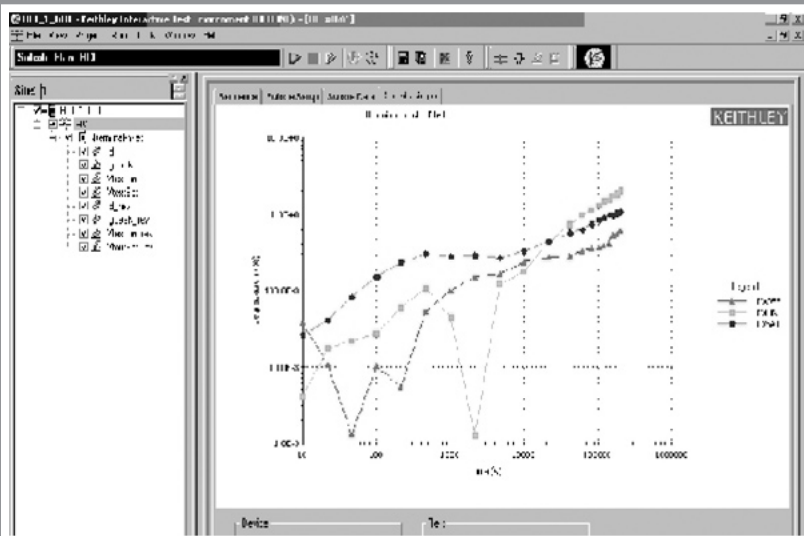


Figure 5. Model 4200-SCS HCI test screen with real-time data plots.

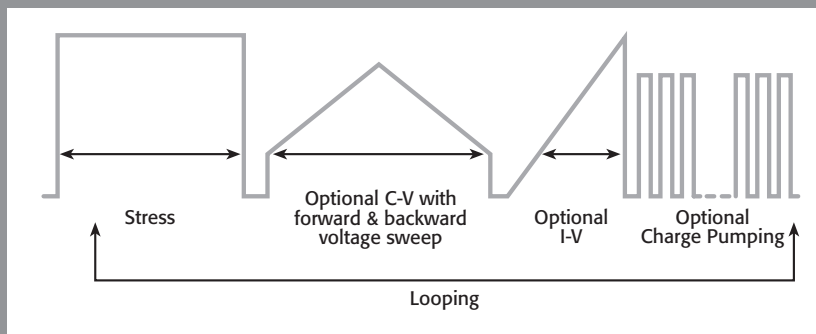


Figure 6. Customized reliability test routine incorporating different stress-measure protocols.

Another reliability system Keithley provides is the Model 4500-MTS Modular Test System. This system, built on a PCI platform, is a parallel device reliability test station for large volume testing. Its design concept is a dedicated SMU for each device under test (DUT). Since each DUT has its own SMU, the test engineer has complete freedom in assigning stress and measure conditions for each test cycle. This is especially valuable when a large sample is needed. It also eliminates the need for a switch matrix, because each DUT has dedicated system resources. A fully loaded 4500-MTS can contain up to 36 independent SMU channels.

Conclusion

Evolving design scales and new materials are making wafer level reliability testing more critical than ever. This is also driving the demand for reliability testing and modeling much further upstream—especially into the R&D process. Keithley has responded with new reliability test tools that are faster, more sensitive, and highly flexible to help drive down the cost of testing and shorten the time to market. ■

Making charge-pumping measurements with the Model 4200-SCS Semiconductor Characterization System

Overview of the charge-pumping technique

Charge-pumping measurements are widely used to characterize interface state densities in MOSFET devices. Recently, with the development of high dielectric (high κ) gate materials, charge pumping has proven especially useful in characterizing charge-trapping phenomena in high κ thin gate films. In thin gate films, leakage current is relatively high due to quantum mechanical tunneling of carriers through the gate. As a result, the traditional technique for extracting interface trap density—collecting simultaneous quasistatic and high frequency C-V measurement data and comparing the difference—can't be used because quasistatic C-V is very hard to achieve at the leakage current level. However, charge-pumping measurements can still be used to extract interface trap density, and the effect of gate leakage can be compensated for by measuring charge-pumping current at lower frequency and subtracting it from measurement results at higher frequencies [1, 2].

The basic charge-pumping technique involves measuring the substrate current while applying voltage pulses of fixed amplitude, rise time, fall time, and frequency to the gate of the transistor, with the source, drain, and body tied to ground. The pulse can be applied with a fixed amplitude, voltage base sweep or a fixed base, variable amplitude sweep.

In a voltage base sweep, the amplitude and period (width) of the pulse are fixed while sweeping the pulse base voltage (**Figure 1a**). At each base voltage, body current can be measured and plotted against base voltage. The interface trap density (D_{it}) can be extracted as a function of bandbending, based on this equation:

$$D_{it} = \frac{I_{cp}}{qAf\Delta E}$$

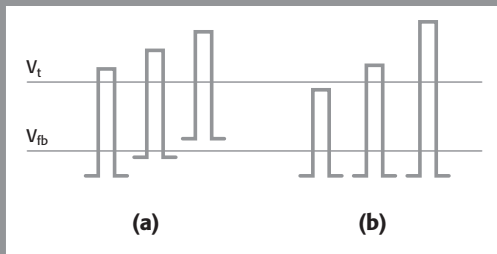
where I_{cp} is the measured charge-pumping current, q is the fundamental electronic charge, A is the area, f is the frequency, and ΔE is the difference between the inversion Fermi level and the accumulation Fermi level [3].

A fixed base, variable amplitude sweep has a fixed base voltage and pulse frequency with step changes in voltage amplitude (**Figure 1b**). The information obtained is similar

Figure 1. Overview of charge-pumping measurements:

(a) Pulse waveform for base voltage sweep; pulse amplitude is constant.

(b) Pulse waveform for amplitude sweep; base voltage is constant.



to that extracted from a voltage base sweep. These measurements can also be performed at different frequencies to obtain a frequency response for the interface traps.

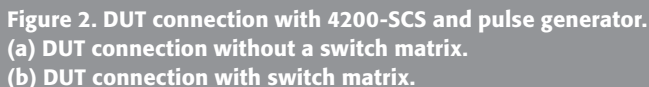
Hardware setup

It's relatively easy to perform charge-pumping measurements and data analysis using a Model 4200-SCS (Semiconductor Characterization System) in combination with a pulse generator (such as Agilent's Model 8110, 81110, or 8112). The KTE Interactive software that runs the Model 4200-SCS can simultaneously control the system's internal Source-Measure Units (SMUs) and external instruments via GPIB with simple C programming. Refer to the Model 4200-SCS Reference Manual and Keithley Application Notes for guidance on using the Model 4200-SCS and KTE Interactive software. **Figure 2a** illustrates the connections for a device under test (DUT) with one of the Model 4200-SCS's SMUs and a pulse generator without a switch matrix; in **Figure 2b**, a semiconductor switch matrix is included in the configuration. This application note describes how to perform charge-pumping measurements with the Model 4200-SCS and an Agilent Model 8110 pulse generator.

Installing the charge-pumping measurement driver

A driver for making charge-pumping measurements with the Model 4200-SCS can be obtained from Keithley Instruments. Currently, this driver supports Agilent Model 8112 and Series 8110/81110 pulse generators. Be sure to request the driver written specifically for the pulse generator model to be used. The driver contains modules that can do base-sweep and amplitude-sweep. To install the driver on the 4200-SCS, copy the source files (.c files) to a new temp folder. Then, go to the **Windows menu bar**, select **Run** and type *cmd*. In the DOS prompt window that pops up, change the directory to the temp folder that contains the source files, and type in *kultcopy ChargePumping* and *enter*. This will install the driver on the Model 4200-SCS.

Note: The driver is currently unsupported. Please exercise caution in using it.



The first thing to do is to enter the proper pulse generator model number (and switch matrix, if necessary) into KCON, the software interface that controls the Model 4200-SCS's internal hardware (SMUs and preamps) and external instruments with GPIB communication. To access KCON, just double-click on **KCON** on the Model 4200-SCS desktop. Then, from the **Tools** menu, select **Add external Instruments > Pulse Generator > HP8110/81110**. **Figure 3** shows a KCON window with this pulse generator added. If the pulse generator model to be used is not included on the supported list, it must be added as a "General Purpose Instrument" (GPI) rather than a "Pulse Generator" (PGU) in KCON. After the pulse generator is added, KCON assigns an instrument ID string to the pulse generator. This ID string could be *PGUx* or *GPIx*, depending on how the pulse generator is added (Pulse Generator Unit or General Purpose Instrument), and *x* could be any number from 1 to 4. This ID will be used as an input parameter for the charge-pumping measurement. It tells KITE software which instrument is on the GPIB and its address.

Figure 3. KCON setup window.

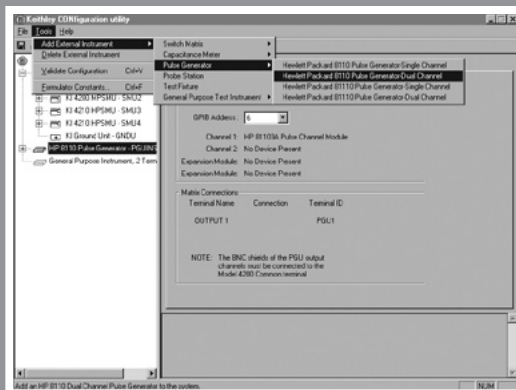
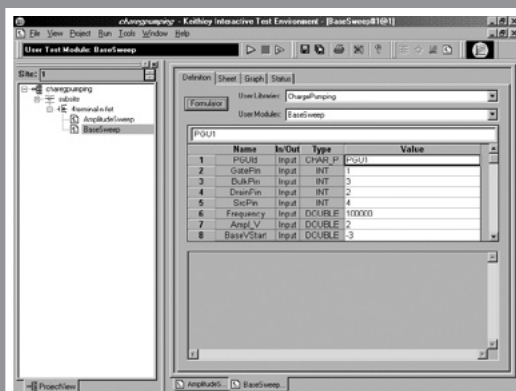


Figure 4. KITE project window.



Setting up a project in KITE

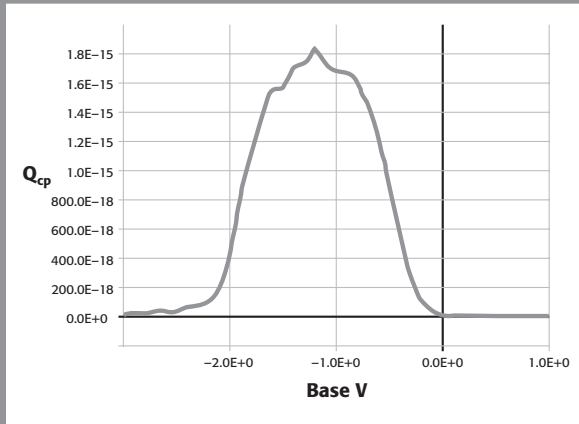
KITE is the main software interface that controls internal hardware and external instruments with GPIB interfaces. Charge-pumping measurement will be performed in KITE interface with a UTM (User Test Module). Refer to the Model 4200-SCS Reference Manual for more details on KITE operation and UTMs.

Double-click on the **KITE** icon on the Model 4200-SCS desktop to bring up the KITE interface. From the **menu bar**, select **File > New project** and type in the project name. Then, go once more to the **menu bar** and select **Project > Make new subsite plan** and type in a valid subsite name. Next, go to the **menu bar** and select **Project > Make**

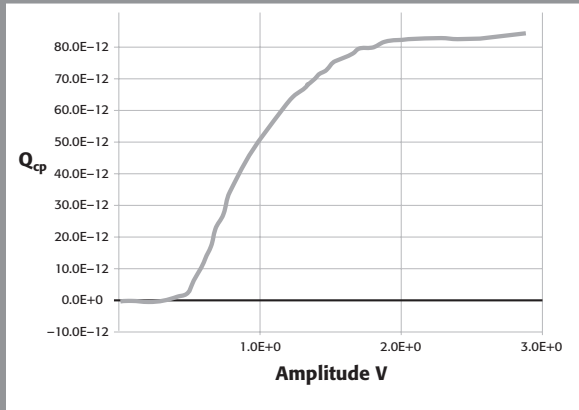
new device plan and choose a device from the MOSFET folder. Finally, go to the **menu bar** and select **Project > Make new User Test Module** and type in the module name, (for example, *BaseSweep*) and click **OK**. Double-click the **BaseSweep** module in the project tree and a setup window will appear. In the setup window, choose user library **ChargePumping** from the drop-down library list, and choose module **BaseSweep** from the User Module list. After the module is selected, a parameter window will appear, as in **Figure 4**. The first parameter on the list is *PGUI**d*, which refers to the ID in KCON. If HP8110 is used and properly configured in KCON, *PGUI* should be used. If HP8112 is used, then *GPLx* should be used, where *x* is a number assigned in KCON. Fill in the parameter list with the proper setup parameters, such as switch connection (if

Figure 5.
Example plot.

(a) Base voltage sweep.



(b) Amplitude sweep.



no switch is presented, then fill in 0 in the *GatePin*, *BulkPin*, *SourcePin* and *DrainPin* field), frequency, pulse amplitude, and duty cycle. After all the input parameters have been properly defined, click the **Save** button (the small floppy disk icon) and the test is ready to run. Click the **Run** button (green triangle icon) to execute the test. The data can be plotted in graph form once the measurement is complete. **Figure 5** illustrates two examples of these graphs.

Calculating D_{IT}

The Model 4200-SCS's Formulator function supports calculating D_{IT} . To activate the Formulator window, click the Formulator button on the definition tab of a test setup window. Enter the formula for D_{IT} as shown in **Figure 6**. The resulting D_{IT} value can be plotted with the system's graphing tools.

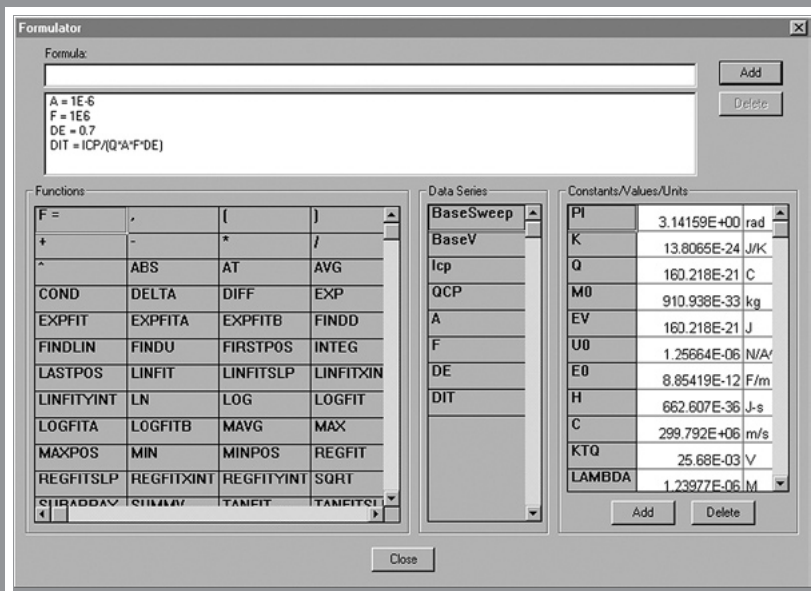


Figure 6. Entering formulas in the Formulator.

Conclusion

The KTE Interactive software on the Model 4200-SCS makes it very easy to make charge-pumping measurements with a pulse generator. With the current driver, one does not need any programming work to do charge-pumping with an Agilent 8112 or 8110/81110 pulse generator. Simple data analysis can be done in the built-in Formulator

and plotted with the powerful graphing tools. This makes the 4200-SCS an ideal tool for characterizing interface properties of gate dielectrics, especially in the area of high dielectric material development. ■

References

- [1] P. Masson, et al., "On the Tunneling Component of Charge Pumping Current in Ultrathin Gate Oxide MOSFETs," *IEEE Elect. Dev. Lett.*, Vol. 20, No. 2, pp. 92-94, 1999.
- [2] S.S. Chung, et al., "A Novel and Direct Determination of the Interface Traps in Sub-100nm CMOS Devices with Direct Tunneling Regime (12~16Å) Gate Oxide," *2002 VLSI Tech. Digest of Tech. Papers*.
- [3] G. Groeseneken, H.E. Maes, N. Beltran, and R.F. De Keersmaecker, "A Reliable Approach to Charge-Pumping Measurements in MOS Transistors," *IEEE Trans. Electron. Dev.*, Vol. ED-31, pp. 42-53, 1984.

Appendix 1: Example source code for base sweep using 8110

```
#include "keithley.h"

int BaseSweep( char *PGUId, int GatePin, int BulkPin, int DrainPin, int
SrcPin, double Frequency, double Ampl_V, double BaseVStart, double
BaseVStop, double BaseVStep, double RiseTime, double FallTime, double
DutyCycle, double LoadImp, double *BaseV, int BaseVSize, double *Icp, int
IcpSize, double *Qcp, int QcpSize )
{
/* USRLIB MODULE CODE */
int index;
int NumPoints;
int PGU1;
char CommandString[50];
int fcnstat;
char TempBuf[20] = "";
int GPIBAddress;
char GPIBAddressStr[10];
double idummy;
int temp = 1;
int PLC;

getinstid(PGUId, &PGU1); // get PGU id from KCON

if (PGU1 < 0)
    return( INVAL_INST_ID ); // No such PGU

getinstattr(PGU1, "GPIBADDR", GPIBAddressStr);
GPIBAddress = atoi( GPIBAddressStr );

if (PguInit(GPIBAddress)< 0) return -1; // Initialize PGU
// set up PGU
if (PguSetup(GPIBAddress, RiseTime, FallTime,DutyCycle, Frequency,
LoadImp)<0) return -1;

//Validate Input parameters
if (GatePin > 72) return(INVAL_PARAM); // Validate pins
```

```

if (BulkPin > 72) return(INVAL_PARAM); // Validate pins
if (SrcPin > 72) return(INVAL_PARAM); // Validate pins
if (DrainPin > 72) return(INVAL_PARAM); // Validate pins
if (Frequency == 0) return(INVAL_PARAM); // Validate pins
if (BaseVStep == 0) return(INVAL_PARAM); // Validate pins

//Initialize return arrays
for (index = 0; index < IcpSize; index ++)
{
    BaseV[index] = DBL_NAN;
    Icp[index] = DBL_NAN;
    Qcp[index] = DBL_NAN;
}

// setup switch matrix if necessary
if(GatePin > 0)
    conpin(PGU1,GatePin, 0);
if(BulkPin > 0)
    conpin(SMU1,BulkPin, 0);
if(SrcPin > 0)
    conpin(SMU2,SrcPin, 0);
if(DrainPin > 0)
    conpin(SMU2,DrainPin, 0);

//Setup SMU
forcev(SMU2, 0);
forcev(SMU1,0);
lorangei(SMU1,1e-10);
if (Frequency >= 1e6)
    PLC = 1;
else if (Frequency >= 1e5)
    PLC = 2;
else if (Frequency >= 1e4)
    PLC = 5;
else
    PLC= 10;

//Initialize current range
setmode(SMU1,KI_INTGPLC,PLC);
limiti(SMU1,1e-2);
measi(SMU1, &idummy);

NumPoints =(int) fabs((BaseVStart - BaseVStop) / BaseVStep) + 1;

// Turn on output on Pulse Generator
sprintf(CommandString, ":OUTPUT1 ON\n");
fcnstat = kbsnd(GPIBAddress, -1, GPIBTIMO, strlen(CommandString),
CommandString );
if (fcnstat > 0) return(GPIB_ERROR_OCCURED);

//Main sweep loop
for (index = 0; index < NumPoints; index++)
{
    BaseV[index] = BaseVStart + index * BaseVStep;
    // Program the pulse height. This one is tricky...so get it right!
    if ( Ampl_V > 0)
    {
        sprintf( CommandString, ":VOLT1:LOW %9.3e\n",BaseV[index]);
        fcnstat = kbsnd(GPIBAddress, -1, GPIBTIMO, strlen(CommandString),
CommandString );
        sprintf( CommandString, ":VOLT1:HIGHIGH %9.3eV\n", Ampl_V+BaseV[index]);
    }
}

```

```

        fcnstat = kibsnd(GPIBAddress, -1, GPIBTIMO, strlen(CommandString),
CommandString );

    }
    else {
        sprintf( CommandString, ":VOLT1:LOW %9.3eV\n", Ampl_
V+BaseV[index]);
        fcnstat = kibsnd(GPIBAddress, -1, GPIBTIMO, strlen(CommandString),
CommandString );
        sprintf( CommandString, ":VOLT1:HIGH %9.3eV\n",BaseV[index]);
        fcnstat = kibsnd(GPIBAddress, -1, GPIBTIMO, strlen(CommandString),
CommandString );
        sprintf( CommandString, ":OUTP1:POL INV\n");
        fcnstat = kibsnd(GPIBAddress, -1, GPIBTIMO, strlen(CommandString),
CommandString );
    }
    if (fcnstat > 0) return(GPIB_ERROR_OCCURED);

    // Measure substrate current
    if (index == 0) delay(200);
    measi(SMU1, &idummy);
    delay(10);
    intgi(SMU1, &Icp[index]);
    Qcp[index] = Icp[index]/Frequency;
}
//turn off output
sprintf(CommandString, ":OUTPUT1 OFF\n");
fcnstat = kibsnd(GPIBAddress, -1, GPIBTIMO, strlen(CommandString),
CommandString );
if (fcnstat > 0) return(GPIB_ERROR_OCCURED);

return OK;
/* USRLIB MODULE END */
}
/* End BaseSweep.c */

```

High throughput gate dielectric reliability testing: Digging out from the backlog

New semiconductor materials and shrinking device dimensions have made reliability testing increasingly important because the gate dielectric behavior of modern devices isn't well understood. Market pressures are pushing devices built of the new high κ materials into production sooner, leaving less time for device characterization, process development, and process integration. Typically, stress-switch-measure testing has been the mainstay of NBTI (Negative Bias Temperature Instability) and TDDB (Time Dependent Dielectric Breakdown) testing. However, the stress-switch-measure configuration doesn't provide the high throughput required to keep pace with today's shorter development cycles. A new approach, using dedicated, cost-effective source/measure channels, provides statistically significant quantities of data in as little as one-third the time of a typical stress-switch-measure system and at a similar price. These dedicated source/measure channels aren't the high cost SMUs (Source-Measure Units) with wide measurement ranges that are typically used, but slimmed-down, targeted source/measure channels.

The reliability of silicon-based devices has become increasingly important as IC manufacturers introduce new dielectric materials and push design and fabrication boundaries to increase performance. In addition to device geometries that are shrinking to <90nm, these new dielectric materials are increasingly important, because smaller device geometries are no longer sufficient to drive product design and roadmaps [1, 2, 3]. Unfortunately, high κ materials aren't as well understood or as easily fabricated as the SiO_2 traditionally used, increasing the need for reliability testing.

For many years, chips had lifetimes of more than 20 years, a period much longer than the targeted design lifetime for the final product, providing reassurance that the IC would continue to perform over time. However, as IC designs incorporate smaller features and new materials, device lifetimes are approaching the lifetimes of the end products, requiring additional testing and modeling to ensure sufficient understanding and confidence in the IC design and production processes.

The primary challenges for gate dielectric reliability testing are device relaxation and high throughput. Device relaxation, a concern for NBTI, occurs soon after the stress is removed from the device, preventing an accurate measurement of the stressed part [4]. This immediate relaxation is a problem for both determining device lifetime on existing technologies and developing models for new technologies and combinations of materials. In addition to device-level degradation, NBTI has performance implications

at the product or chip level [5, 6]. For all reliability testing, high throughput is required to minimize the amount of time required to gather statistically significant quantities of data for lifetime prediction and modeling. Testing many parts simultaneously provides statistically significant amounts of data in the same time as testing one part [7].

The traditional wafer-level solution uses a small number of full-featured (wide dynamic range) SMUs (Source-Measure Units) and a switch to perform the stress/measure cycling. This solution has advantages for high resolution measurements, but doesn't provide the high throughput required for today's fast-changing IC development environment. Fortunately, however, there is a cost-effective, high channel count solution that provides the high throughput and seamless stress-measure transition to minimize device relaxation; this solution will be described later in this article.

Traditional stress-switch-measure method

The most popular test system for wafer-level reliability testing in technology research and development labs consists of four to eight high performance SMUs, a switch matrix, and control software, paired with a prober and a hot chuck (**Figure 1**). This shared SMU configuration is adequate and widely used for reliability testing where limited stress conditions are acceptable. The strength of this system is the high quality data it can produce for device modeling and intensive study of small numbers of devices. In addition to reliability testing, this type of system has comprehensive source and measure ranges to provide complete DC device characterization, which is useful for development, design verification, and failure analysis. See **Figure 2** for a measurement connection diagram.

Mainframe

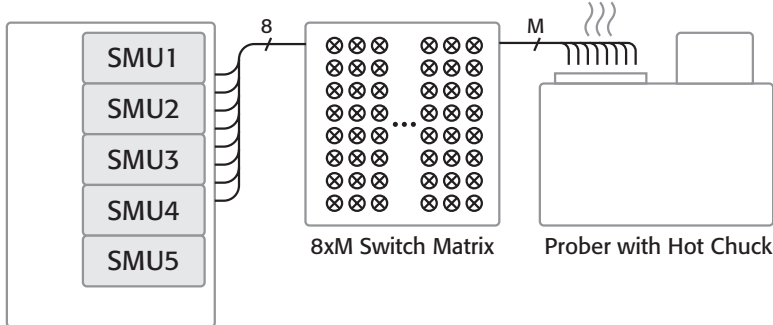
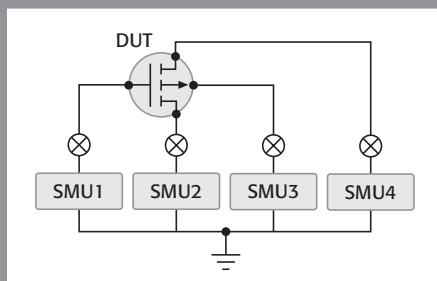


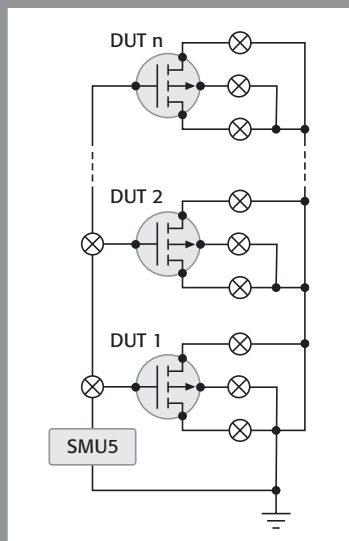
Figure 1. Diagram of traditional stress-switch-measure, shared SMU, test system. SMUs 1–4 are dedicated to measurements, while SMU 5 provides the stress voltage for all DUTs. See Figure 2 for DUT connection details.

Figure 2. Configuration using Shared SMU (stress-switch-measure) system in the two test configurations:

- (a) Measure: DUT connections during a measure interval.
- (b) Stress: DUT connections during a stress interval, sharing one SMU to stress all DUTs.



(a)



(b)

Switching permits multiple DUTs (devices under test) to share the relatively expensive SMUs. For reliability testing, the switching is used to move each DUT from the measure setup (**Figure 2a**), where multiple SMUs are used to characterize the device, to the stress condition (**Figure 2b**), where all DUTs are held at the appropriate stress voltage by a single SMU. See “Concerns related to switch-centric systems.”

Shared measure method

A second type of system uses a modified switching configuration. Instead of using shared Source/Measure Units (SMUs), where the source and measure are coupled, the source and measure functions are separated. This permits the use of many sources, which allows for unique stressing, while switching shares the measure capability across all DUTs. This second approach, shared measure, provides unique voltage stressing across DUTs, but doesn't fully address the throughput, because the measurement capability is still shared across many DUTs. Also, the measurement capability is typically less sensitive than in the shared SMU arrangement. Sharing measurements prevents detailed monitoring during the stress, because readings during stress still have to be “walked in sequence” to each DUT via switching. This approach is commonly seen in package-level reliability testing.

After reviewing the drawbacks listed, it's clear the traditional shared SMU method (Stress-Switch-Measure) is good for high resolution measurements and characterization,

Concerns related to switch-centric systems

- Switching causes delays between the stress and measure steps (**Figure 3**). Switching delay allows the device to relax or recover, potentially eliminating the behavior to be measured.
- Measurements performed sequentially across the DUTs (**Figure 3**).
 - Missing a key failure event: As the measurement is being performed on one DUT, a key failure event occurring on another DUT could be missed.
 - Some systems provide no measurements or infrequent measurements during the stress interval.
 - As the number of DUTs increases, so does the chance of missing a failure event.
 - As the number of DUTs increases, shorter stress times are difficult or impossible to achieve, because subsequent stress times are largely determined by the number of switch+measure cycles required to test previous DUTs.
- Each DUT undergoes different stress times (**Figure 3**).
 - Requires more complex software to track the unique stress/measure cycle times for each DUT.
 - Convolved data analysis and correlation across DUTs.
- Potential voltage “glitch” during hot switching could damage lower voltage devices.
- Limited number of stress conditions available; the number of unique stress voltages is equal to the number of SMUs available for stressing.
- All DUTs are stressed in a parallel configuration, so DUTs that fail in a low resistance (high current) condition must be removed from the test pool quickly, to minimize undesirable test conditions, such as a sagging stress voltage.

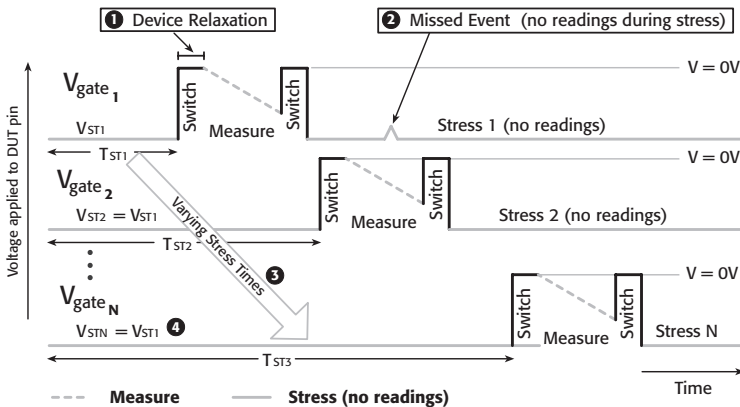


Figure 3. Example stress/measure cycle for a stress-switch-measure approach.

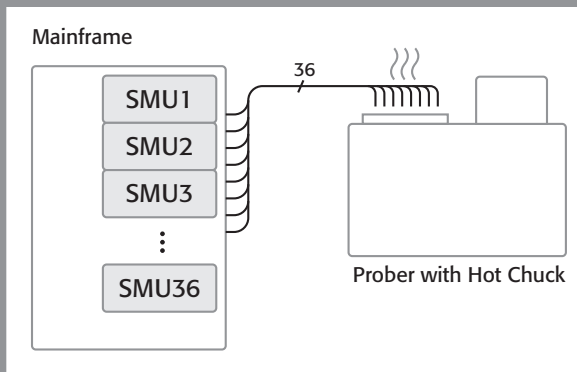
Note: ❶ device relaxation; ❷ missed event during stress on DUT #1; ❸ different stress times for each DUT; ❹ single V_{stress} voltage shared for all DUTs (i.e., $V_{ST1} = V_{ST2} = \dots = V_{STN}$). This is a simplified diagram showing the stress and measure on one pin per DUT; typical stress/measure configurations will have non-zero voltages on 2–3 pins of the 4 pins on a typical FET DUT.

but has drawbacks related to throughput, stress-to-measure transition time, and visibility into device performance during stress. The shared measurement approach takes a step towards addressing some of the drawbacks of the shared SMU arrangement, but doesn't provide the throughput necessary to supply statistically significant data in shorter test times.

High throughput parallel reliability test using dedicated SMU per DUT

Addressing the issues relating to throughput, relaxation, and measure during stress requires a new source/measure architecture. A third type of system combines the strengths of the two commonly used methods described previously (**Figure 4**). The primary feature of this new, high throughput parallel test solution is that each DUT signal-of-interest has a dedicated SMU to source voltage and measure current. One way to accomplish this configuration would be simply to increase the number of wide-dynamic-range, high cost SMUs capable of femtoamp measurements. This isn't acceptable, because the cost and rack space requirements of the wide-dynamic-range SMUs would be prohibitive.

Figure 4. Diagram showing dedicated source/measure per pin for NBTI/TDDB testing.



Therefore, this new system doesn't just use a higher number of wide dynamic range SMUs, but uses an SMU tailored to NBTI and TDDB testing for ULSI CMOS structures and devices. This tailored SMU provides the measurement ranges appropriate for NBTI and TDDB testing, reducing costs, and permitting a larger number of SMUs for the same cost as the wide-dynamic-range SMUs. This new system can provide up to 36 SMUs to test up to 18 transistors (**Figure 6**). For example, a 36-channel (18-transistor) capability could be provided for the same price as a system with five wide-range SMUs and an appropriate switch.

Summary of Dedicated SMU Test Architecture for NBTI/Tddb

- Dedicated SMU for each DUT signal
 - Eliminates switching, allowing a seamless transition between stress and measure (Figure 5)
 - Eliminates SMU sharing (Figure 6)
 - Independent source value for each signal
 - Continuous monitoring of gate/drain currents during stress interval
 - DUT failures don't disrupt the stress levels of the remaining DUTs
 - Characterize multiple DUTs at the same time
 - Get statistical quantities of data quicker
 - Test multiple technologies simultaneously
 - Reduce time for process development
 - Up to 36 unique stress conditions for quicker lifetime extraction
 - Independent A/D converter for each channel
 - Provides frequent measurements during the stress interval
 - Permits increased measurement quality
 - All measurements are time correlated (Figure 5)
 - More accurate measurements during DUT degradation, especially important where failures are related to a gradual signal change or change of the noise in the signal
- 4500 software and hardware architecture optimized for high channel count configurations
 - Eliminates channel coordination issues, as software and hardware overhead is increasingly cumbersome for configurations requiring >8–10 SMU channels
 - Channel grouping permits similar DUTs to easily share common configurations and test sequences

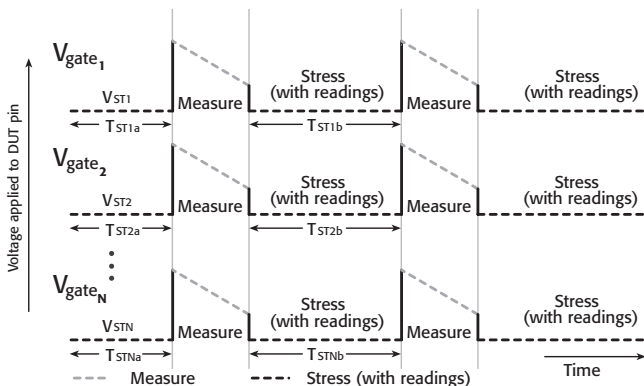


Figure 5. Example stress/measure cycle for dedicated SMU approach. Note that the time for each stress interval is the same, that each voltage stress (V_{ST}) can have a unique value, and that there are readings taken during the stress intervals.

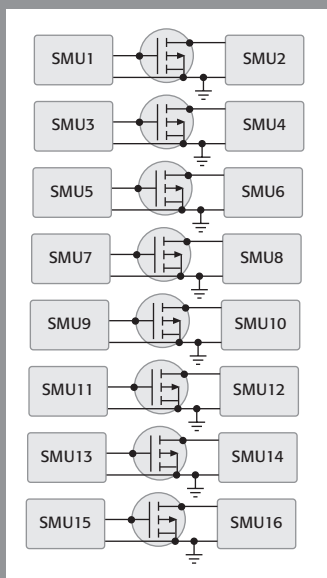


Figure 6. Eight DUTs, with a dedicated SMU for each signal of interest. During stress, the even-numbered SMUs connected to the drains are typically set to 0V.

The use of additional SMUs allows this new system to eliminate switching for the stress/measure cycling common to both NBTI and TDDB. This high channel count, parallel test solution, which provides up to 3× higher throughput for both NBTI and TDDB, provides many advantages. For NBTI, the higher throughput is provided by increasing the number of unique simultaneous stress conditions available (i.e., each DUT can have a unique stress condition). For TDDB, the higher throughput is provided by continuous measurement of each gate current (I_g) with a dedicated A/D converter. For both of these tests, this means statistical data can be collected much faster, providing earlier visibility into the behavior to improve modeling algorithms and reduce time to market.

Dedicated source/measure capability also provides benefits in addition to throughput. For measurements, there is one A/D converter for each signal, which provides better measurements than a switched system, because the eliminated switching time can be used to provide longer integration times and permit

better noise rejection. For NBTI, the dedicated voltage source and A/D converter provides a seamless transition between the stress and measure test intervals to minimize the amount of device relaxation to the single millisecond range for each DUT. Also, frequent measurements can be made during the stress, allowing tracking of the stress effects without interrupting the stress condition. For sourcing, there is no switching, which eliminates switching times, device relaxation, and hot-switch glitches.

Of course, assembling a high throughput system means integrating many source/measure channels, which requires resolving programming, triggering, data retrieval, and rack space issues. The Keithley 4500-MTS addresses all of these multi-channel integration headaches, while providing up to 36 independent source/measure channels in a 5U instrument (**Figure 7**).

Hybrid Solution

The 4500-MTS's SMU per DUT pin solution increases reliability testing throughput, but doesn't always eliminate the need for a wide dynamic range SMU capability. Therefore, a hybrid system, consisting of both wide dynamic range SMUs (e.g., Keithley 4200-SCS) and a 4500-MTS, can be considered. This hybrid approach provides wide dynamic range sourcing and picoamp measurement capabilities for both pre- and post-stress characterization, along with the 4500-MTS's seamless transitions and stress monitoring capability.



Figure 7. 4500-MTS and 4510-QIVC (Quad I/V Card).

The 4500-MTS mainframe can support up to nine cards; each card contains four stress-measure channels. Each channel is optimized for NBTI and TDDDB testing of ULSI CMOS structures, providing both appropriate source voltages and current measure ranges to provide a cost-effective solution. In addition, the 4500-MTS uses a standard PCI architecture running Windows®. The 4500 cards are controlled through a driver, which are accessed via a program written in any modern development environment: National Instruments' LabVIEW® or LabWindows®/CVI, Microsoft's Visual Basic® or Visual C++®, or another development environment that can access a 32-bit Dynamic Link Library (DLL).

Conclusion

New materials and accelerated development cycles are putting increased pressure on reliability testing. To address these new requirements, Keithley has developed a new test architecture to address the need for more reliability test data. The 4500-MTS provides high throughput testing of ULSI CMOS for both NBTI and TDDDB to supply statistical quantities of data in as little as one-third the time of an existing shared SMU system. This data is available not only during the typical measure intervals, but also during stress, providing greater insight into device behavior during stress. Finally, switching during the stress/measure cycle is eliminated, providing a seamless transition during the test and virtually eliminating device relaxation. ■

NBTI - Negative Bias Temperature Instability

NBTI is a phenomenon where change in the gate-channel interface causes degradation in pMOS device performance. The degradation is typically tracked as the increase of the transistor threshold voltage (V_T) and degradation of the drain current (I_D). This degradation can reduce yield through failures during burn-in or in the field [5, 6]. NBTI testing doesn't have a well-defined industry standard, although most parameters and methods are similar across the industry [3].

The test for NBTI involves a stress/measure cycle (**Figure 8**), alternating between stress, V_G on the gate, and measure, where a V_G - I_D sweep, or single-point measurement of I_D , is performed. Throughout the test, the device is held at an elevated temperature to decrease test times, which can range from a few hours to a few weeks. While stressing the gate, the other transistor pins are grounded (**Figure 6**). The V_G - I_D sweep or other test is performed to determine the shift in device performance, typically the shift in the threshold voltage (V_T). To minimize the amount of data collected, measurement intervals are typically made on a logarithmic time basis, resulting in total stress interval times of 1s, 10s, 100s, etc.

To differentiate between the current readings made during the measure intervals and current readings taken during the stress interval, the terminology used here is “measurements” for the measure interval and “stress monitor” or “stress reading” for the stress interval. The 4500-MTS has the capability of monitoring all gate currents (I_G) during the stress interval.

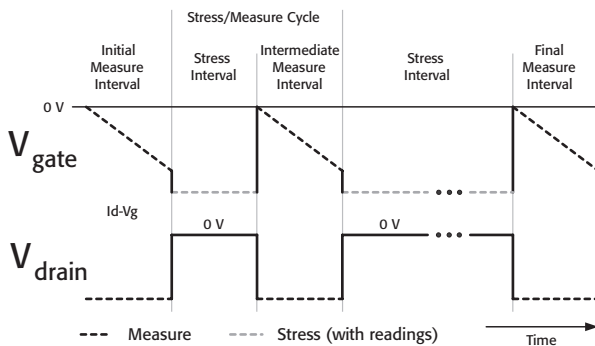


Figure 8. Typical pMOS NBTI Stress/Measure Diagram (switching effects not shown).

TDDB – Time Dependent Dielectric Breakdown

TDDB tests the lifetime of the gate dielectric via voltage stressing over time. Unlike NBTI, which is only concerned with device degradation, TDDB includes both degradation and failure of the dielectric. Unlike NBTI, TDDB has a standard for testing, JESD 92 [9], published by JEDEC. However, this standard has yet to be fully adopted by the industry.

The test method for TDDB can be considered a subset of NBTI. TDDB has similar pre- and post-stress measurement intervals to determine device characteristics. Unlike the multiple stress/measure cycles of NBTI, TDDB consists of a single stress interval that isn't typically interrupted by any measure intervals (**Figure 9**). During stress, just the gate current (I_g) is monitored. New materials and smaller dimensions are changing the type of failure from a hard breakdown to a soft breakdown [8]. A hard breakdown is an abrupt change in I_g . A soft breakdown is seen as an increase in the noise of the signal (I_g), along with a gradual increase in I_g over time, possibly tens of seconds. Throughout the test, the device is held at an elevated temperature to decrease test time. While the gate is being stressed, the other transistor pins are grounded.

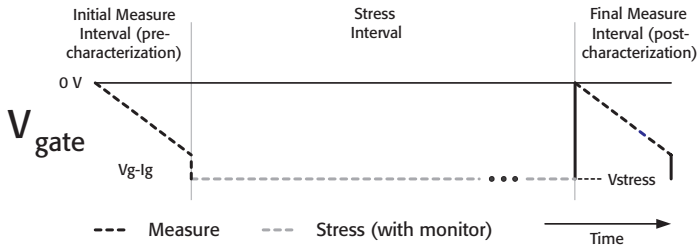


Figure 9. Typical TDDB Stress/Measure Diagram for Constant Voltage Stress (CVS) method (switching effects not shown).

References

1. P. Clark. (2004, May). "Scaling died at 130-nm so innovate, says IBM CTO," *Silicon Strategies*. [Online]. Available: www.siliconstrategies.com/article/printableArticle.jhtml?articleID=19400132
2. International Technology Roadmap for Semiconductors (ITRS) 2003 Roadmap. [Online] Available: <http://public.itrs.net/Files/2003ITRS/Home2003.htm>.
3. D. Schroder and J. Babcock, "Negative bias temperature instability: Road to cross in deep submicron silicon semiconductor manufacturing," *J. Appl. Phys.*, vol. 94, pp. 1–18, July 1, 2003.
4. S. Rangan, N. Mielke, and E. Yeh, "Universal Recovery Behavior of Negative Bias Temperature Instability," in *Proc. IEEE Intl. Electron Devices Mtg.*, p. 341-342, Dec. 2003.
5. V. Reddy, et al. "Impact of Negative Bias Temperature Instability on Product Parametric Drift," in *Proc. IEEE ITC Intl. Test Conf.*, Oct. 2004, pp. 146-155.
6. Y.H. Lee, et al. "Effect of pMOST Bias Temperature Instability on Circuit Reliability Performance," in *Proc. Intl. Electron Devices Mtg.*, 2003, pp. 353-356.
7. Jeff Kuo, et al. "Reducing parametric test costs with faster, smarter parallel test techniques," Keithley Instruments White Paper, 2004.
8. J.W. McPherson, J Kim, A. Shanware, H. Mogul, J. Rodriguez, "Trends in the ultimate breakdown strength of high dielectric-constant materials," *IEEE Trans. Elec. Dev.*, vol. 50, no. 8, pp. 1771–1778, Aug. 2003.
9. "Procedure for Characterizing Time-Dependent Dielectric Breakdown of Ultra-Thin Gate Dielectrics," JEDEC JESD92, Aug. 2003.

Improved thermal stability of copper vias using a cyclical stress test

Via stress migration (VSM) occurs when the relaxation of thermal stresses in copper interconnects causes defects to form under a via [1]. Tests for this phenomenon usually involve measurements of the change in via resistance after an extended bake of 100 to 1000 hours at fixed temperature. The long cycle time for VSM testing delays the feedback loop for process optimization. Isothermal VSM tests also depend on the choice of optimum stress temperature. An alternative VSM test is proposed that cycles temperature through the region of largest creep rate. A 12-hour cyclic test was found to give larger via resistance shifts than a 50-hour isothermal test. Further improvements in the VSM test were obtained by tracking smaller resistance shifts to improve the statistics of the failure rate and reduce wafer-wafer variations. Detection of small via resistance shifts required an electrical test that could measure shifts of less than 0.002Ω without excessive current. The improved via stress migration test has been applied to copper/low κ structures to identify a new type of failure mode associated with resist poisoning.

Introduction

Stress migration involves a balance between diffusion and thermally induced stress in metal interconnects. At low temperatures, the metal is under high tensile stress from differential thermal contraction relative to the dielectric. As temperature is increased, the stress relaxation rate from diffusion increases but the absolute stress decreases. A model has been developed by McPherson and Dun [2] for the stress induced creep of a metal line embedded in a dielectric. An important parameter in the model is the temperature (T_0) at which the metal transitions from tensile to compressive stress. The creep rate from this model is shown in **Figure 1**. It is a strong function of test temperature and T_0 .

The parameter T_0 can depend on the thermal history of the copper and the initial high temperature stress state of the copper.[3] Therefore, the choice of an appropriate test temperature is critical to quantify VSM reliability. Typical test conditions reported in the literature are 168–500 hours [1, 4] at a fixed temperature of 150°C to 300°C. The long anneal time extends the feedback for VSM tests to weeks. A technique that is faster and independent of test temperature would be invaluable to speed the development cycle.

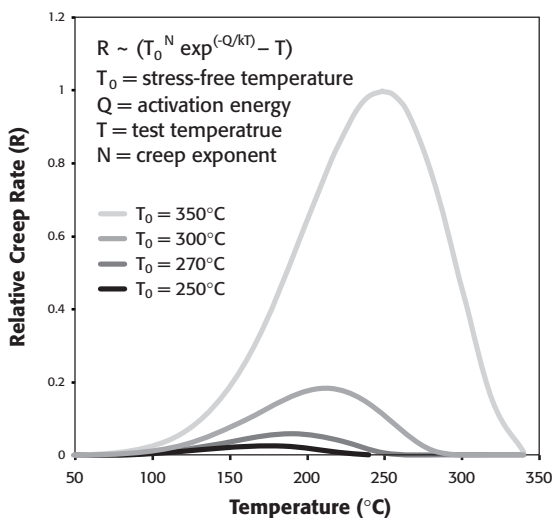


Figure 1. Creep rate and test temperature for a range of stress/temperature cross over points.

Experiment

In contrast to single temperature tests, a cyclic annealing method can include the full temperature range of high creep rate independent of process condition or thermal history of the copper. The critical temperature range for stress migration is 150°C to 250°C , depending on the stress state of the copper. A furnace was used to slowly cycle temperature 4 times from 150°C to 250°C over a period of 12 hours. The heat/cool cycle was dependent on the furnace and was approximately 1 hour for heating and 2 hours for the cooling.

The resistance of 120 0.2mm Kelvin vias was measured across a wafer before and after the thermal treatment. A Keithley Instruments S600 series parametric tester and a fully automatic prober was used for the measurements. The typical via resistance was 0.6Ω and the desired precision for measuring shifts in resistance was 1% or 0.006Ω . This level of precision is a very strict requirement for a parametric tester. The measurement issue is further compounded by limits on the current level per via. Currents larger than $\sim 1\text{mA}/\text{cm}^2$ risk heating the via and/or damaging it. Requirements for a stress migration test used here were $0.0025\Omega/1\sigma$ repeatability before and after the thermal stress and less than 0.5mA of current. **Figure 2** shows the repeatability for an optimized Kelvin via resistance test. Repeated measurements of 30 Kelvin via sites were measured

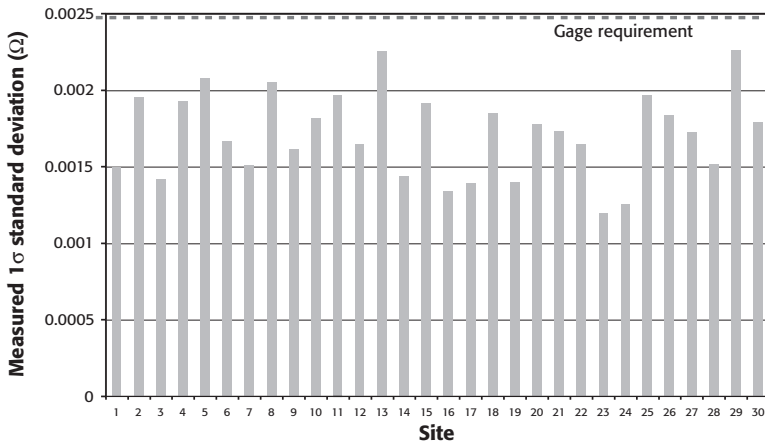


Figure 2. Repeatability of a 30-site/30-day repeatability test for measurements of Kelvin vias with nominal resistance of 0.6Ω .

over a 30-day period. We find that the repeatability of this Kelvin test is typically within $0.002\Omega/1\sigma$. This level of repeatability permits accurate measurement of resistance shifts less than 1% in a via stress migration measurement.

Using a smaller resistance shift to classify failed vias increases the number of failures and therefore improves the statistics of the test. A single failure criterion helps to compare data from different process variations rapidly and to monitor reliability performance. **Figure 3** shows that adopting a small percent shift ($<10\%$) reduces the wafer-to-wafer spread of the failure rate because of the increase in number of failed vias. However, the choice of failure criterion can affect conclusions from VSM tests. Limiting the analysis to large resistance shifts ($>10\%$) assumes that only large defects contribute to the failure rate, but will reduce the number of failed vias. Using a 2% shift as a failure criterion will improve statistics, but might include the wrong types of failures. Failure analysis of vias with large and small resistance shifts is shown in **Figure 4**. The nucleation point of the void is almost always at the lower corner of the via where stress concentration is largest. Once a void nucleates, the driving mechanism for growth is usually the same. Therefore, detection of small resistance shifts early in the VSM process will give an accurate measurement of nucleation density of voids that may eventually grow into open failures.

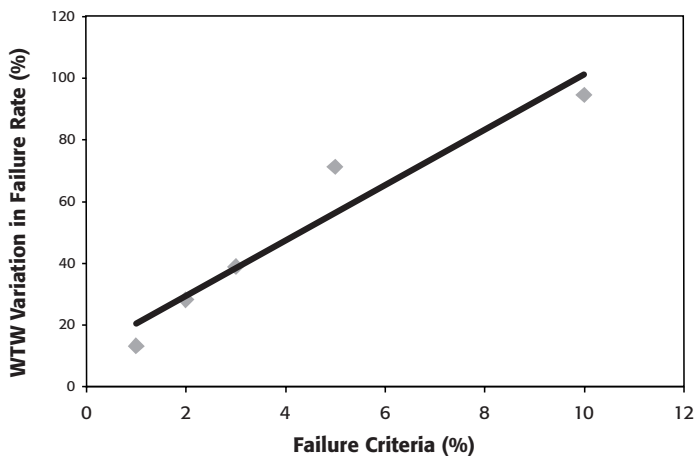
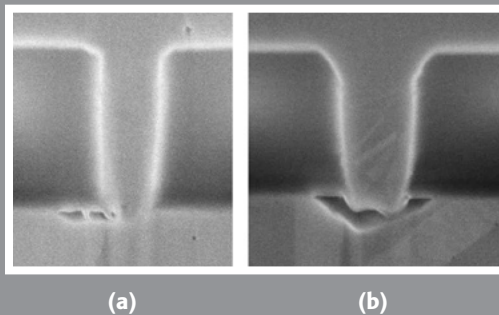


Figure 3. Wafer-to-wafer variation in failure rate as a function of failure criteria for a set of eight wafers with the same process.

Figure 4. Voids under via after 150°C–250°C cyclic thermal test. Resistance shifts were (a) 1% and (b) 10%. Voids are similar to that observed with static thermal tests.



Results and Discussion

We find that a cyclic test method is able to achieve significant shifts in via resistance after only 12 hours of anneal, whereas a fixed temperature anneal at 200°C shows smaller resistance shifts, even after 50 hours of anneal as shown in **Figure 5**. Typical VSM voids after the cyclic test are similar to voids observed after static tests, as shown in **Figure 4**. The larger resistance shifts with a cyclical test could be a reflection of a non-ideal choice of stress temperature for the isothermal test or that the cyclical test is more

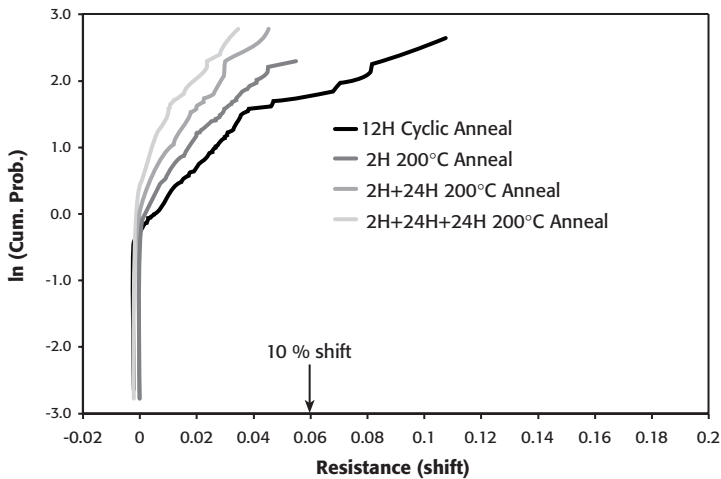
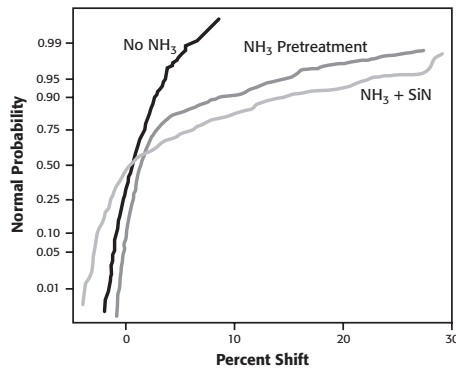


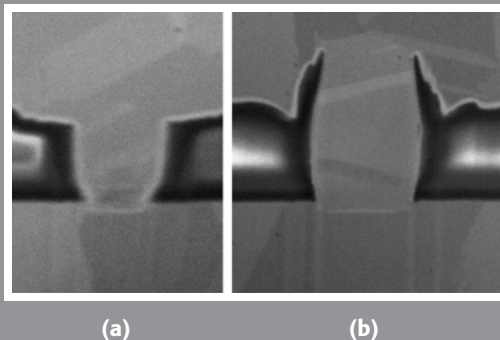
Figure 5. Comparison of a cyclic anneal (150°C–250°C cycle repeated four times over 12 hours) to a fixed temperature anneal at 200°C. Each curve is the sum of three wafers for each test condition.

Figure 6. VSM shift for three dielectric diffusion barrier processes with a PECVD low κ dielectric and isolated Kelvin vias. Each curve is the sum of three wafers.



effective at creating voids due to the “ratcheting” effect of tensile/compressive stress. Although the cyclic anneal cannot be used to extrapolate lifetime, it does allow for rapid and quantitative screening of stress migration performance.

Figure 7. Single Kelvin via profile for the data shown in Figure 4 with no ammonia (a) and ammonia in the bulk film and pretreatment (b). Fence formation around the via is due to poisoned resist near the via.



Another advantage of the cyclical test is that it is less sensitive to the thermal history of the copper. If the low temperature stress state of the copper changes with time (through relaxation), then the zero-stress transition temperature (T_0) will change. This will have a dramatic effect on the creep rate at a fixed temperature, as shown in **Figure 1**.

The test procedure described here is sensitive to other process related defects, in addition to via voids. **Figure 6** shows the resistance shift of an isolated Kelvin via with copper, PECVD low κ dielectric, and three different dielectric diffusion barrier processes. We find that a barrier process that contains excessive amounts of ammonia has significantly worse VSM relative to an ammonia free process. Failure analysis showed that the non-optimized barrier process caused poisoning of the photoresist in the region of the via and the formation of “rings” around the via as shown in **Figure 7**. A poor via profile leads to less thermal stability of the via and ultimately a reliability concern. The stress migration test described here was able to detect the presence of problems with the via profile when the via resistance alone did not look anomalous. Subsequent optimization of the ammonia/low κ process was able to improve the via profile and via stress migration performance was improved.

Conclusion

In conclusion, we have shown that a 12-hour cyclic thermal treatment is as effective as a 50-hour isothermal treatment for inducing resistance shifts in a via stress migration test. These tests require very high precision from the test equipment. We have shown that 0.6Ω Kelvin vias can be measured with a repeatability of better than $0.0025\Omega/1\sigma$ using a low current test. The greater precision of this test allows improved statistics for summarizing via stress migration data by using a smaller failure criterion. We have applied this test to copper/low κ structures and have identified an additional failure mode associated with via-fence formation. ■

Acknowledgements

We would like to thank T. Mountsier, M. Sanganeria, R. Shaviv, D. Vitkavage, R. Havemann, and M. Kollrack for useful discussions and assistance in processing the wafers.

References

- [1] E.T. Ogawa, J.W. McPherson, et al., *Proc. Int'l Rel. Phys. Symp.*, 2002, p. 312.
- [2] J.W. McPherson and C.F. Dunn, *J. Vac. Sci. & Tech.*, B5(5), 1987, p. 1 321.
- [3] S.H. Ree, Y. Du, and P.S. Ho, *Proc. Int. Interconnect Tech. Conf.*, 2001, p. 89–92.
- [4] K. Doong, R. Wang, et al., *Proc. Int'l Rel. Phys. Symp.*, 2003, p. 156.

Reducing parametric test costs with faster, smarter parallel test techniques

Introduction

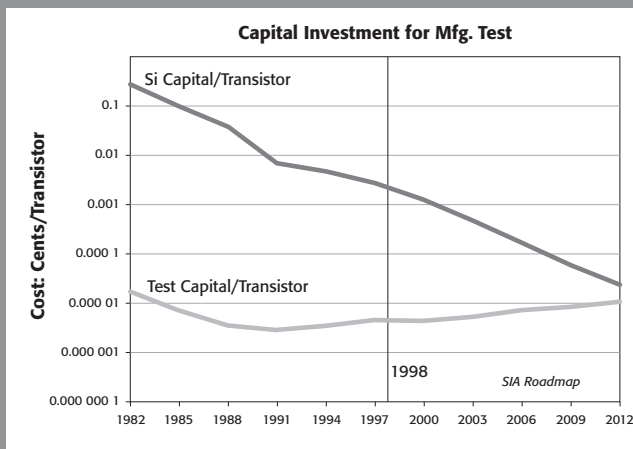
The 1999 SIA roadmap included an ominous prediction—that by about 2012, the cost of test (COT) per transistor would approach the cost of fabrication per transistor, as indicated in **Figure 1** [1].

Although the validity of extrapolating a decade into the future in an industry that delivers a new technology generation every two to three years is debatable, the trend line is nonetheless alarming, particularly for a manufacturing step that, while vitally important, is too often undervalued. Typical strategies for decreasing cost of test include testing less, testing more efficiently, testing differently, and reducing the cost of the testers used [2].

Implications for parametric cost of test

Although a wafer's overall cost of test is dominated by the cost of functional testing, many fabs use a common organizational and reporting structure for both parametric and functional testing. This common structure sometimes colors perceptions of the economics associated with parametric test. However, the economics of parametric test differ significantly from those associated with functional test:

Figure 1.
Fabrication
and test cost
trends.



- Parametric test uses a sampling strategy, rather than measuring every die on every wafer.
- Parametric test results are used for process control and yield improvement, not for binning finished integrated circuits (ICs).
- Depending on the supplier, equipment in the parametric test cell can often be reused extensively—in fact, a recent analysis indicates it's possible to achieve up to 85% capital equipment reuse over five or more process nodes.
- Parametric test involves measuring a wide array of signal types, ranging from femtoamp DC leakage to 40GHz RF s-parameters.

Applying a typical cost of ownership model to parametric test as used in volume production, then performing sensitivity analysis, reveals that while a 50% decrease in initial capital equipment cost decreases the cost of test per wafer by only 15%, a 50% test time reduction (TTR) delivers nearly a 50% decrease in cost of test per wafer. Ongoing TTR is dependent on choosing a system with a robust, flexible software and hardware architecture that allows performing field upgrades cost-effectively. These upgrades make possible an on-going entitlement to TTR (and therefore, COT reduction) in the range of 10% per year. The high value associated with TTR has led us to focus this work on the "test more efficiently" strategy for reducing parametric COT.

More efficient parametric test with parallel test

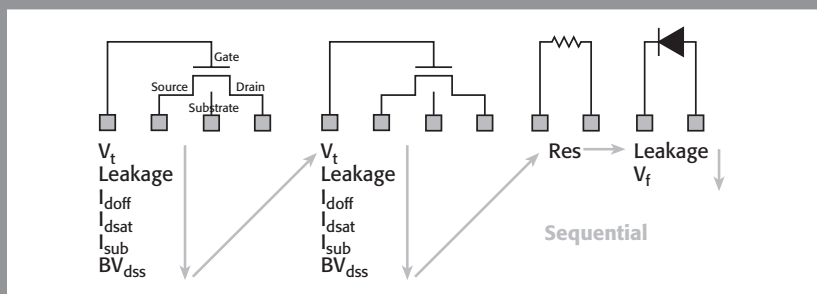
Although parallel testing is a well-accepted technique for TTR in functional test, it has only recently become available on parametric testers. This is due in part to the complexity of managing the wide range of measurements involved in parametric test, which requires measuring the electrical parameters of the key devices that form the basic building blocks of all integrated circuits: resistors, capacitors, diodes, transistors, inductors, varactors, etc. Measurements are performed on specially designed test structures, typically located in the scribe lines of product wafers. Force-measure sequences can be programmed for single bias points or as multiple bias points swept in time. For the DC portion of the test suite, a typical set of requirements for measurement resources such as source-measure units (SMUs) might resemble those outlined in **Table 1**.

A schematic example of sequential mode testing of the devices within a test site might look like **Figure 2**.

A modern parametric test system can have up to eight identical high resolution and high power SMUs. That means that in a sequential test mode, when a resistor is being measured (requiring one SMU), then up to seven SMUs are sitting idle. By measuring multiple mixed types of devices simultaneously within a single probe touchdown and

Table 1. Example SMU requirements.

Device	Type of Test	Measurement level/ method	Number of SMUs needed
CMOS transistor	Threshold voltage (V_t)	Max. g_m of 20 steps	3-4
	Leakage	1pA	2
	I_{doff}	100fA	2
	Saturation current (I_{dsat})	μA to mA	2
	I_{sub}	nA to 10 μA , sweep 50 steps	3
	Drain-source breakdown (BV_{dss})	V	1
Resistor	Resistance	μA	1
Diode	Leakage	pA	1
	Forward junction voltage (V_f)	Force single current	1

**Figure 2. Example of sequential mode testing of devices within a single test site.**

thereby increasing utilization of both the tester and the prober, parallel test delivers higher throughput. For example, two resistors, one diode, and one transistor could possibly be measured simultaneously by independently and asynchronously performing different connect-force-measure sequences on all four devices at the same time (**Figure 3**).

Figure 4 illustrates a schematic example of parallel mode testing of the devices within a test site that maximizes instrumentation resources.

A parametric test system must have several fundamental hardware and software capabilities to support parallel test:

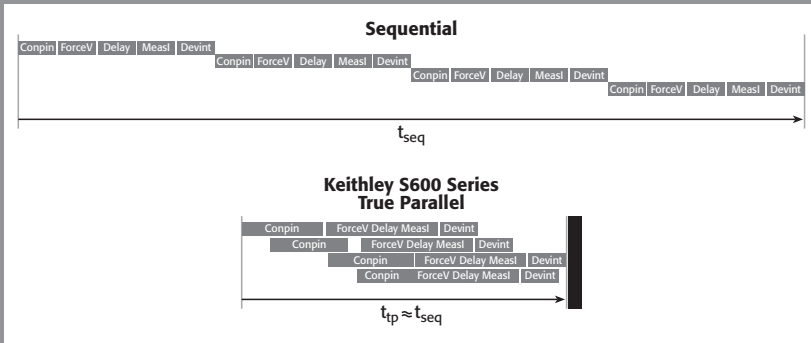


Figure 3.

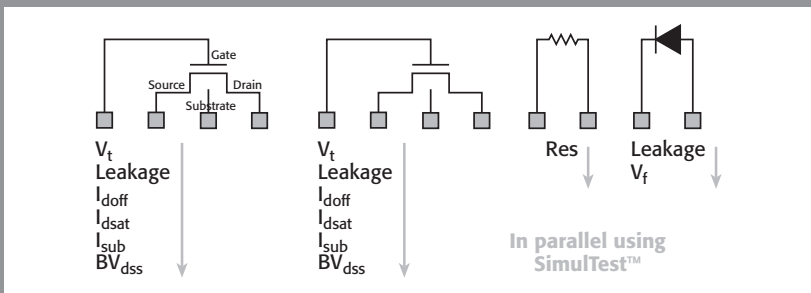


Figure 4.

- Identical, redundant measurement paths, all with lab-grade resolution to support fast test setup and high code reuse, and minimize contention between hardware resources.
- Source and measurement hardware able to run independently. For example, each piece of hardware needs to have its own high precision A/D converters, as well as its own embedded real-time logic processors and communications channels.
- The test execution environment must be multi-threaded.

Test structure design is another important consideration when attempting to get the maximum benefit of parametric parallel test. Depending on the company or fab's philosophy, test structures are typically designed to optimize one of these aspects: minimiz-

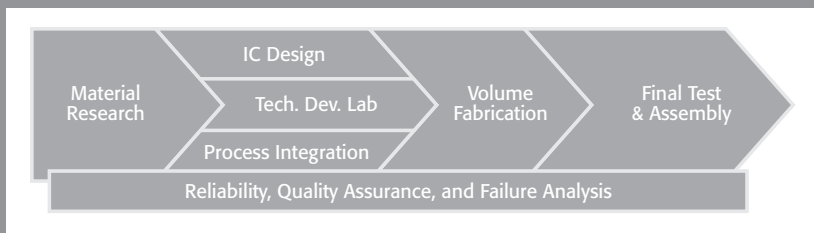


Figure 5.

ing wafer area (lots of shared pads) and/or maximizing the quality of test results (little or no sharing of critical pads). Structures designed to maximize the quality of results typically allow a higher degree of parallel testing. Fabs whose philosophy for sequential test structures leads them to minimize area typically achieve appreciable throughput improvement initially, then can realize additional TTR by making small changes to their test structure designs as they intercept future mask changes.

Parametric test is used primarily during the process development, process integration, and volume fabrication portions of an IC's lifecycle (shown in **Figure 5**).

Throughout the process ramp, the amount of parametric testing performed per wafer ranges widely and constantly, with roughly 100 times more parameters measured during process integration (when the learning curve is steepest) than during volume fabrication (when throughput and incremental yield improvement are more important). Therefore, parametric parallel test offers complementary benefits:

1. The ability to acquire the same amount of data in significantly less time during volume fabrication.
2. The ability to acquire more data in the same amount of test time during process development.

Volume production parallel parametric test—same amount of data in less time

One volume production logic IC manufacturer performs 300 parametric tests per site on the usual variety of device types. Fast integration (17ms) for signal averaging is used, and the fab's philosophy dictates optimizing test structures for high data integrity—the test devices share few probe contact pads and the scribe line test insert isn't optimized for minimum area. This case allows a high degree of parallelism using existing test structures and probe cards, and the fab achieved 1.7 times higher throughput in measurements at the sites overall, not including prober indexing (wafer movement) time between sites (**Table 2**).

Table 2.

Test Mode	Test Time per Site
Sequential Test	98s
Parallel Test	56s
Test Time Reduction	42%
Throughput Improvement	1.7×

Process development parallel parametric test—more data in the same amount of time

The less obvious benefit of parametric parallel test is the ability to acquire more data in the same amount of time. This use case can occur, for example, during process development, when the learning curve on new materials and devices is the steepest and the opportunity to shorten time-to-market is the greatest. Time-to-market is a primary profitability metric for any IC product. During process development, fabs need to obtain an enormous amount of data quickly for statistical analysis to determine process sensitivity and variability, for verification of process and device models, and for performing corner testing to produce initial process control limits.

Voltage-ramped breakdown (VRB) is one of the reliability tests used during process development to characterize gate capacitors and inter-level dielectrics (ILDs). The test is a very common check of damage to the gate dielectric from a poorly formed oxide or from damage induced by processing. In the case of inter-level dielectrics and the copper damascene process, this test is an important indicator of the integrity of the copper diffusion barrier layer and capping layer interface. The growing use of low κ dielectrics makes this test even more important because of their lower intrinsic breakdown fields

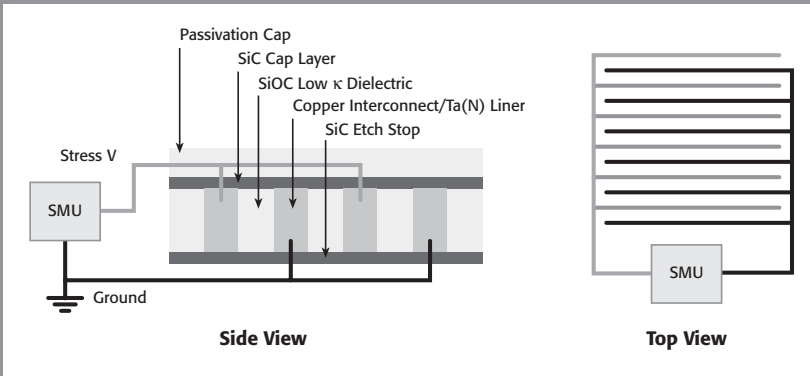


Figure 6.

and lower interfacial adhesion strengths. The typical test structure for inter-level dielectric reliability (**Figure 6**) in a copper/low κ process is an inter-digitated metal-dielectric comb structure comprised of two parallel metal lines with dielectric between the lines.

VRB is a destructive test where voltage across the dielectric is ramped from 0V to as high as 100V, and leakage current is monitored. An abrupt increase in the measured leakage current from one voltage bias point to the next indicates the dielectric has catastrophically broken down and the voltage immediately before breakdown occurred is recorded. Due to the statistical nature of the failure mechanisms, many die are measured across the wafer, and cumulative probability of breakdown voltages is compared between different processes. Test time depends more on the voltage at which the dielectric fails (with good devices taking longer to test) and less on whether multiple DUTs are tested in parallel. A typical ramp rate for the VRB test of comb capacitors with low κ dielectric might be 1V/s. This ramp rate is slow relative to other breakdown tests because the voltages can be quite high and the leakage currents can be transient for low κ dielectrics. If breakdown occurs at 5MV/cm with a $0.2\mu\text{m}$ dielectric spacing, then it would take 100s to get to the 100V breakdown voltage. Faster ramp rates would result in even higher breakdown voltages because the effective time at a voltage is shorter. Such high breakdown voltages might exceed the voltage limit of the tester or change the failure mechanism in the low κ comb structure.

Given the relatively long time needed to perform this test, the number of die tested must be limited to obtain reasonable test throughput. A standard parametric test sampling strategy might be to measure only 16 out of the 121 die available on the wafer, and only one of the 12 structures available within a die (**Figure 7**).

This strategy was believed to provide an optimal tradeoff between test time and amount of data. Process effects such as dielectric erosion and other phenomena were

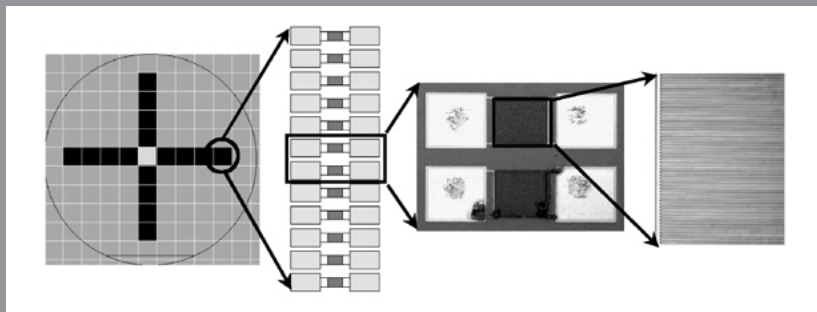
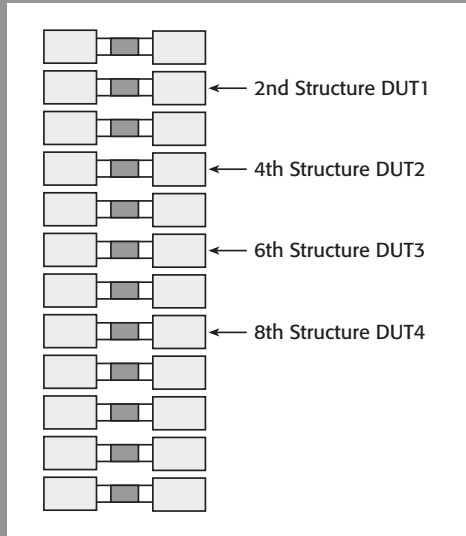


Figure 7.

Figure 8.



always observed to occur on spatial scales consistent with the chosen die sampling that spanned the wafer, so it was not believed that measuring more structures in closer proximity (more than one device per die) would provide any additional process information.

The test time for 16 die was approximately one hour. Because it was modeled that measuring four DUT in parallel within the same die would not increase the test time and in the interest of discovering new processing phenomena, three more DUTs (#2, 3, 4) were measured in each site (**Figure 8**).

Figure 9 is the resulting Cumulative Probability Plot (CPP) of VRB test results from the test wafer.

When only one DUT per die was tested, the median breakdown field was $\sim 4\text{MV/cm}$ and the distribution was a very broad Gaussian, with no sign of multimode failures. Based on this data set, one might conclude the integrity of the low κ dielectric layer was compromised across the whole wafer, so the wafer and the process it represents should be rejected. However, the curve for the four DUTs combined that was acquired in nominally the same test time shows that the median breakdown field was 50% higher at 6MV/cm , and the distribution appears to be bimodal. A bimodal distribution indicates there might be a very local process issue affecting the integrity of the low κ dielectric, but the general integrity of the low κ dielectric is good. This conclusion is significantly

Figure 9.

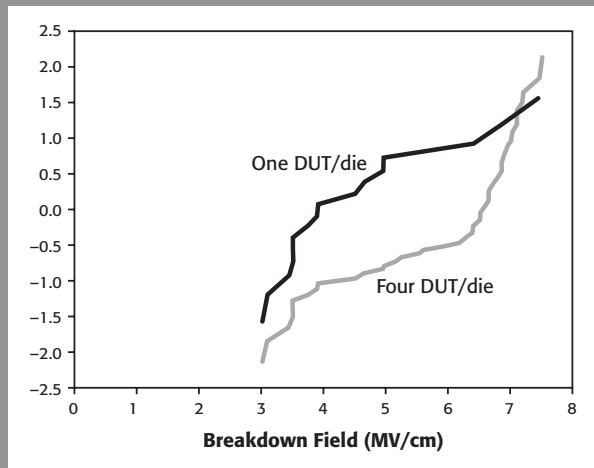
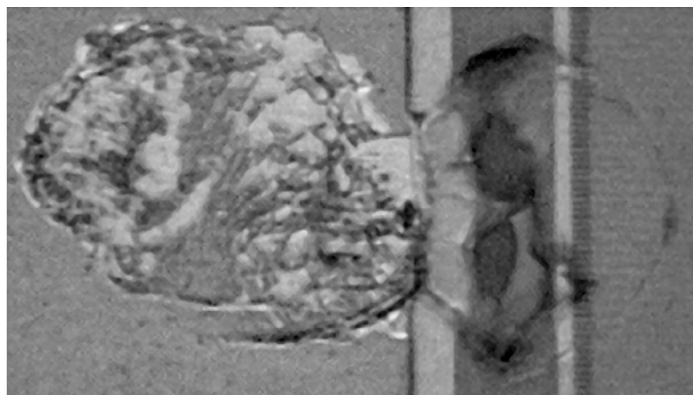


Figure 10.



different than the one drawn from the one-DUT-per-die curve. Failure analysis of the failed die showed localized cracking of the dielectric passivation layer in the neighborhood of the die during test (**Figure 10**).

This cracking of the passivation layer was sufficient to degrade the breakdown properties of the low κ dielectric.

Summary and conclusions

The advent of parallel testing capability takes modern parametric test systems out of the extrapolation that predicts wafer test cost will exceed wafer fabrication cost in three technology nodes. Parallel parametric test has been shown to deliver the same data in substantially less test time in the volume production use case. It was also shown to deliver substantially more data (and learning) in the same test time during process development, with the discovery of a new Cu/low κ process integration effect. By coordinating the development of test structures for parallel test with scheduled mask changes, parallel parametric test also provides ongoing opportunities for decreasing parametric cost of test in volume production. ■

References

1. S. Sengupta, et al., "Defect-Based Test: A Key Enabler for Successful Migration to Structural Test," *Intel Technology Journal*, 1st Quarter 1999, <http://www.intel.com/technology/itj/q11999/articles/art_6c.htm> (29 Mar. 2004).
2. S. Carlson, "ATE Struggles to Keep Pace with VLSI," *EE Times*, December 13, 2001, <<http://www.us.design-reuse.com/articles/article2278.html>> (29 Mar. 2004).

OVERCOMING THE MEASUREMENT CHALLENGES
OF ADVANCED SEMICONDUCTOR TECHNOLOGIES

SECTION V

Femtoamp DC Leakage for Mobile ICs

Tips, tricks, and traps for advanced SMU DC measurements

An SMU (Source-Measure Unit or SourceMeter® instrument) will normally give an accurate measurement, but sometimes errors can creep in, and special methods are needed to overcome them. This article shows how to do some advanced measurements that may require more features than are normally used. The measurement problems covered are common to all SMU users, but some of the solutions are unique to Keithley SMUs.

Understanding an SMU

An SMU is actually four instruments in one: a precision voltage source, a precision current source, a voltmeter, and an ammeter. SMUs are used in semiconductor device testing, optoelectronic test, materials research, and even as general lab instruments. In **Figure 1**, the source block represents both the voltage source and current source capability. In reality, an SMU is always acting as both a voltage source and a current source.

The V_{measure} circle represents the built-in voltmeter capability. Note that the voltmeter gives feedback to the source block, which means that it can be used to control it. The I_{measure} circle represents the built-in ammeter, and it, too, can control the source block.

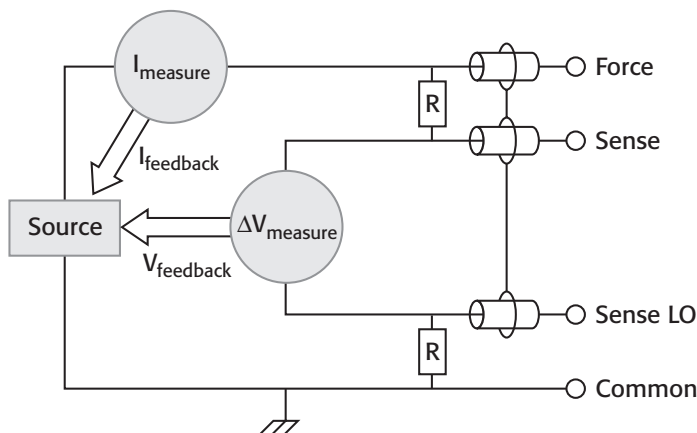


Figure 1. An SMU includes a precision voltage source and a precision current source (shown here as one block), a voltmeter, and an ammeter.

The voltage and current output go between the Force terminal and the Common terminal. By Kirchhoff's loop rule, all current flowing out of the Force terminal must flow into the Common terminal. The return path is normally direct, although in some cases, another SMU can act as the return path.

Note the Sense and Sense LO terminals. These special high impedance terminals are used to sense voltage more accurately at the DUT (device under test), and no current flows in them. They are used only where extreme accuracy is needed or in some special applications.

Interpreting published specifications

It's important to understand both the source and measure specifications, because they have a large impact on many measurements. **Table 1** gives partial specifications for the Keithley Model 4210-SMU found in the 4200-SCS Semiconductor Characterization System. Most of the examples in this article are affected by one or more of these specifications.

Table 1. Typical specifications for the SMUs in the Model 4200-SCS

Range	Compliance	Measure Resolution	Measure Accuracy	Source Resolution	Source Accuracy
20 V	1.05 A	20 μ V	0.01% + 1 mV	500 μ V	0.02% + 1.5 mV
1 A	21 V	1 μ A	0.1% + 200 μ A	50 μ A	0.1% + 350 μ A

Note the 20V specification. The *Compliance* is 1A, which is the maximum current the instrument can supply on that range. The *Measure Resolution* gives the smallest change the voltmeter can detect on the range stated—in this case, 20 μ V, or seven digits of resolution.

Measure Accuracy refers, again, to voltage measurement. The first number, 0.01%, is the *gain* number; multiply it by the reading to get the error. For example, in measuring 20V, the gain accuracy would give a 2mV uncertainty. The second number, offset, normally dominates when making measurements near zero. For example, measuring 0V on this range would give an uncertainty of 1mV.

The *Source Resolution* refers the instrument's ability to act as a voltage source. It's the smallest change that can be sourced out of the Force terminal. Note that the source resolution is 500 μ V, compared to the 20 μ V for measure resolution; as a general rule, an SMU can measure about ten times better than it can source. All sourcing resolution is limited by that 500 μ V resolution. For example, sourcing 2.80005V would not be a problem, but trying to source 2.80004V wouldn't work; it would be rounded to the nearest 500 μ V.

The *Source Accuracy* is related to how close to an exact real voltage can appear at the Force terminal. As with the Measure specification, the first number is called the gain number and is multiplied by the total output voltage. If we wanted to source, for example, 20V, we would have about a 4mV accuracy. The offset number has the greatest impact near zero volts. Setting the output to 0V on the 20V range, for example, could result in as much as 1.5mV at the output.

Looking at **Table 2**, if a user set an SMU to output 3.0005V, the actual output would be somewhere between the minimum and maximum specifications. If the user then set the SMU to output 3.001V, the output voltage would again be between the minimum and maximum specification. But since the step from 3.0005V to 3.001V is within the error band (the maximum specification for 3.0005V is more than the minimum specification for 3.001V), the output, instead of increasing by 500 μ V, could actually decrease by 3.7mV. In reality, this seldom happens, because offset, which creates this inaccuracy, is normally constant from one step to the next.

Table 2. The effect of SMU source accuracy

Desired Output (V)	Minimum Spec. (V)	Maximum Spec. (V)
3.0005	2.9984	3.0026
3.0010	2.9989	3.0031

Source accuracy example: MOSFET transconductance measurement

Figure 2 shows the effect of source accuracy. Here, we use a Model 4200-SCS to determine the transconductance of a MOSFET by sweeping the gate voltage in 2mV steps while measuring the resulting drain current. The transconductance is derived by taking the differential of the drain current with respect to the gate voltage, $dI_{\text{drain}}/dV_{\text{gate}}$. The differential calculation of transconductance is a very good test of an instrument, because it will magnify any errors and noise. Notice the periodic bumps in the transconductance (lower) curve. Where did they come from? **Figure 3** shows the changes in drain current and gate voltage. The plot of changes in drain current contains small bumps, while gate voltage seems to be a perfect 2mV step. Why are there bumps in the drain current when there are none in the gate voltage?

Figure 4 shows the same test again, but with the SMU set to measure the gate voltage. Remember—the SMU can measure about ten times more accurately than it can source. Here, we see 300 μ V bumps in the gate voltage, corresponding to the bumps in the drain current. This is more than 10 \times the source accuracy of the Model 4200-SCS, but by measuring, we were able to detect it. If we now differentiate the drain current with respect to gate voltage, we get the curves in **Figure 5**. With an accurate measurement of source voltage, we get a good and reliable transconductance measurement. To

Figure 2. MOSFET drain current (upper) and transconductance (lower) curves obtained by increasing the gate voltage in 2mV steps. Note the errors in the transconductance.

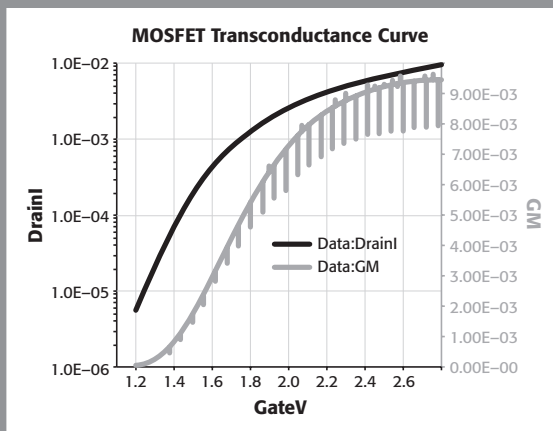
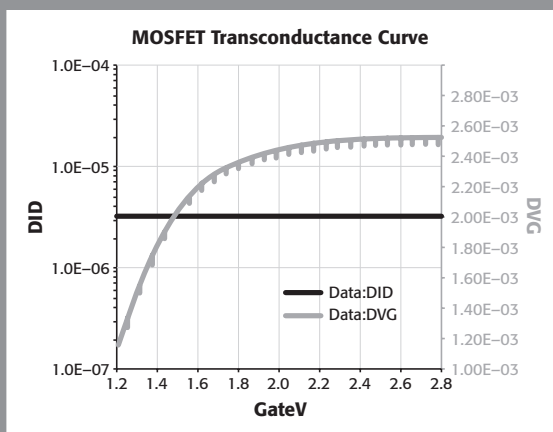


Figure 3. Looking at changes in drain current shows errors.



sum up: When an SMU produces an obviously strange measurement result, review the instrument's specifications closely.

Floating SMU for more source accuracy

Another way to improve the source accuracy of an SMU is to use the Sense LO terminal (**Figure 1**), which acts as a reference point for all voltage measurements and sourcing on the SMU. The SMU floats on top of the Sense LO terminal. This gives several opportunities. If, for example, one put a 1V battery between the Sense LO terminal and

Figure 4. Measuring gate voltage with the SMU's voltmeter function reveals the cause of the problem: 300 μ V errors in the gate voltage.

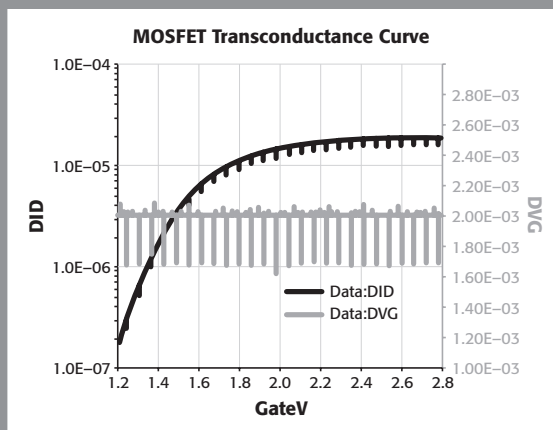
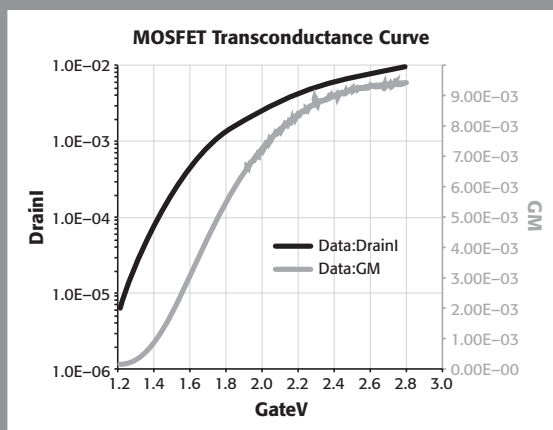


Figure 5. Differentiating the drain current with respect to the gate voltage gives an accurate result.



Common, the entire SMU would float 1V above the Common—which leads to a way to get greater accuracy. Remember, in the previous examples, we were trying to make a 2mV step while the SMU was on a 20V range. We needed the 20V range because the highest voltage used in the application was 3V. If we could use the 2V range, we would have ten times the source resolution—50 μ V—and avoid the 300 μ V error. We do that by raising the Sense LO terminal above the Common—not with a battery, but with another SMU set to 1V (**Figure 6**). Then we set the first SMU to sweep from 0 to 1.8V. The total

Figure 6. Using a second SMU to float the first SMU 1V above ground makes it possible to use the 2V range, cutting the 300mV gate voltage error to 50mV.

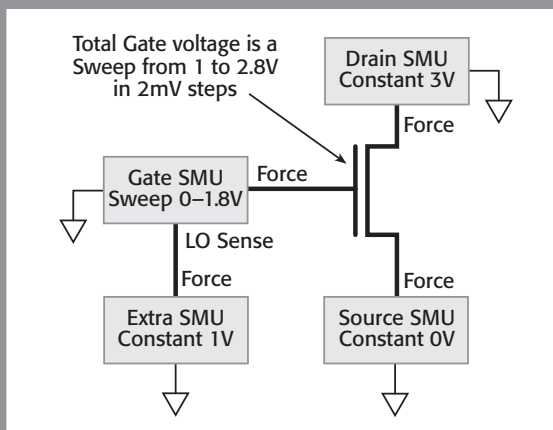
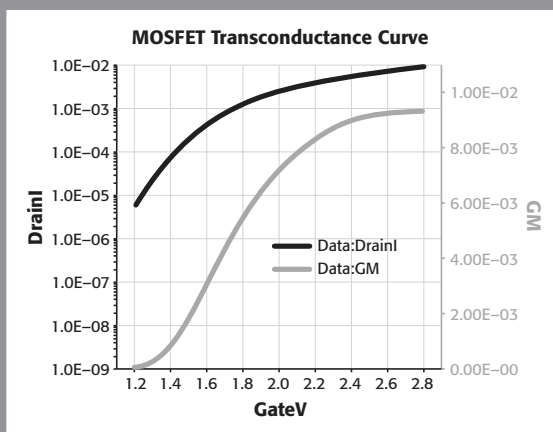


Figure 7. Floating the gate SMU on top of an extra SMU makes it possible to use the 2V range, reducing gate voltage error for error-free results.



sweep is still 1 to 2.8V in 2mV steps, but the gate SMU never needs to produce more than 2V, and the steps are very accurate. The resulting transconductance curves are shown in **Figure 7**.

The SMU as an ideal voltmeter

An ideal voltmeter has infinite input impedance (defined as many orders of magnitude more than the impedance of what's being measured). It draws no current from the circuit being measured, and does not load it down.

To make an SMU into an ideal voltmeter, configure it as a current source, and set the current level to zero. No current can then flow into or out of the terminal, so it has effectively infinite impedance. In practice, the impedance is about one million divided by the full-scale current. For example, if the full scale current were set to $1\mu\text{A}$, the unit's input impedance would be about $1 \times 10^{12}\Omega$. If full scale current were set at 1pA , the input impedance would be about $1 \times 10^{16}\Omega$.

Another key characteristic of an ideal voltmeter is its offset current, which is the current the voltmeter itself generates. The example in **Figure 8** shows how both the input impedance and the offset current can cause a problem. **Figure 8** shows a high impedance source: a 1V source in series with a $1\text{G}\Omega$ resistance. Ideally, the SMU voltmeter should measure exactly 1V . Setting an SMU to act as a voltmeter normally causes it to act as a current source with zero amps output on the 100nA range. For this example,

Figure 8. The input impedance of an SMU configured as a voltmeter is the reciprocal of the current range.

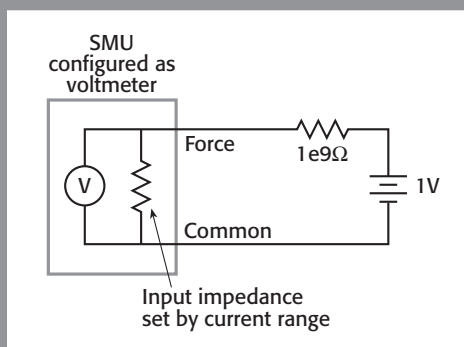
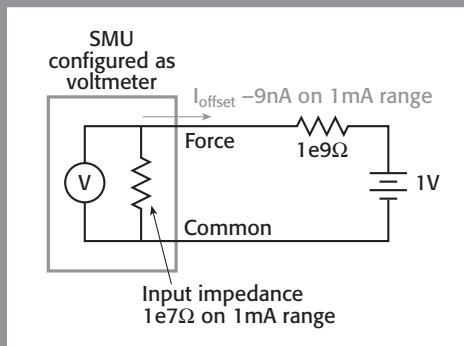


Figure 9. Mis-setting the current range to 1mA gives an input impedance of $10\text{M}\Omega$, which loads down the circuit under test. In addition, it allows a -9nA offset current.



however, we override the normal settings and put the SMU on the 1mA range (**Figure 9**). This gives an input impedance of $10^7\Omega$, which is two orders of magnitude less than the impedance of the source we want to measure and loads down the circuit to give a reading of 10mV instead of 1V.

But there's an even worse problem: the SMU's offset current. On the 1mA range the offset current specification for the Model 4200-SCS is 150nA. We measured it for this experiment, telling the SMU to measure its own offset current, and got -9nA , which is considerably better, but -9nA through a $1\text{G}\Omega$ resistance gives about 8V, so the SMU terminal would attempt to float to 8V. In reality, the SMU floated to -2V , because of the range to which it was set, and clamped there.

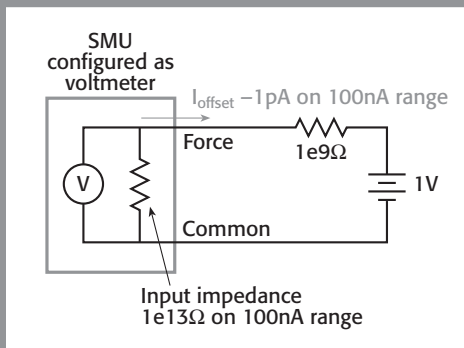
Figure 10 shows the experiment repeated, this time letting the SMU set itself as it normally would—as a current source set to zero amps on the 100nA range. Now the input impedance is four orders of magnitude greater than that of the source, so we should get a good measurement without loading down the circuit. In addition, the offset current on the 100nA range is less than 1pA, so we get an accurate measurement of the battery voltage.

Floating differential voltmeter

An SMU can act as a floating differential voltmeter by making use of the Sense LO terminal. A floating differential voltmeter can be five to ten times more accurate and more sensitive than a standard voltmeter. In addition, it can often take the place of two standard voltmeters and be easier to use as well.

In the previous example, we used the Source and Common terminals. Here, we will use the Sense and Sense LO terminals. **Figure 11** shows a string of three 20Ω resistors, with a 3V battery at the top and a 2V battery at the bottom, so the whole string

Figure 10. Setting the SMU correctly gives $10\text{G}\Omega$ input impedance and -1pA offset current.



is floating 2V above ground. We wish to measure the voltage across the resistors, but not change the current through them. A standard voltmeter, referenced to common, would not be suitable—we need a floating differential voltmeter. We set an SMU to be a voltmeter—a current source set to zero, on the 100nA range. We can set the SMU to the 2V range, whereas if we used a standard voltmeter, we would have to use a 20V or 30V range, which would reduce accuracy and cause problems with offset current, as in the previous example.

A word of warning here: The Sense LO terminal on this particular SMU has about 100k Ω of input impedance, and sometimes that affects the accuracy of the measurement.

Table 3 shows the result of the measurement. Notice it was possible to measure the actual resistor values with great accuracy because the instrument was on the 2V range and thus had an accuracy of better than 150 μ V.

Figure 11. Measuring the voltages across these resistors is best done with a floating differential voltmeter

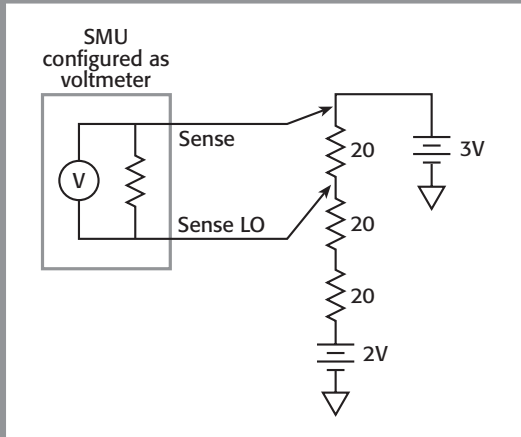


Table 3: Actual node voltages and resistance values measured

Node	Voltage measured	Actual resistor value
Resistor 1	334.04mV	20.04 Ω
Resistor 2	334.75mV	20.08 Ω
Resistor 3	336.40mV	20.18 Ω

The SMU as an ideal ammeter

An ideal ammeter should have zero input impedance—or at least many orders of magnitude less than the circuit being measured. An SMU will act as an ideal ammeter if it is programmed to act as a voltage source set to zero volts. It will then try to hold zero volts at its terminals, and will measure all current that flows in or out. The input impedance of a Model 4200-SCS in this configuration is about one divided by the current range in microamps. For example, on the $1\mu\text{A}$ range, the input impedance is about 1Ω .

An ideal ammeter has low offset voltage, which is voltage that the ammeter itself generates. An SMU can add or subtract voltages to compensate for any offset voltage. In fact, if we use the Sense terminal, we can sense the voltage on the device and compensate for any voltage drop in the probes or test leads. **Figure 12** shows a large MOSFET; we want to measure the characteristics of the drain. We've added a 10Ω resistor in series with the source in order to simulate excessive test probe resistance. This resistor will add an extra voltage burden on the source terminal and allow the transistor to float a little above ground.

We'll connect an SMU as an ammeter to the source terminal and measure the transfer characteristics of the transistor twice—once with the Sense terminal of the SMU disconnected (lower curve in **Figure 13**), and once with it connected (upper curve). The lower curve shows less current flow, because the voltage on the source terminal rises above the common, depending upon how much drain current is flowing, which reduces the gate-to-source voltage. The upper curve, then, is accurate, while the lower one is inaccurate.

Figure 12. Connecting the SMU as shown can compensate for the drop across the source resistor.

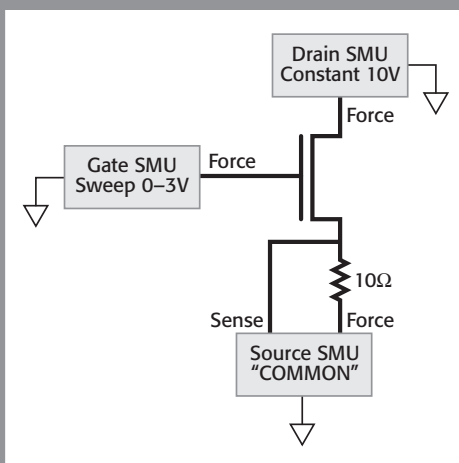
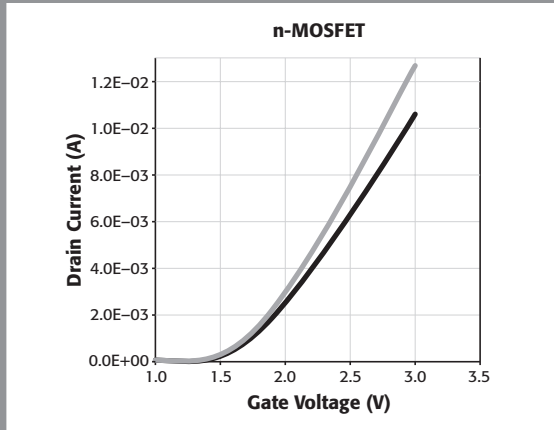


Figure 13. Connecting the SMU's Sense lead as shown in Figure 12 corrects the error.



Conclusions

There are many possible sources of error in precision measurement. Applying the methods outlined here should make it possible to avoid many of them. ■

Parametric test hardware for ultra-low current measurements

Increasingly, parametric testing of MOSFETs in CMOS ICs requires very accurate measurement of various source, drain, and gate currents, and characterization of low current phenomena. The tester, cabling, and chucks used must be properly integrated from the perspective of the device under test and the specific instruments used. The overall approach must include attention to system grounding and shielding, minimization of system noise, consideration of measurement settling errors, and proper attention to measurement speed.

Parametric characterization of MOSFETs typically requires extremely sensitive measurements of the drain current during the off-state. Also, the transition from the on-state to the off-state (i.e., subthreshold current) is crucial to the performance of a CMOS circuit. Typically, off-state current is in femtoamps. Gate leakage current is also crucial in device performance. Furthermore, there are many low current phenomena of particular importance to device reliability, such as gate-induced drain leakage, which must be characterized.

Too often, when a semiconductor characterization system is configured, one concentrates on DC parametric instrumentation while neglecting the rest of the system. For ultra-low current measurements, however, it is very important to have a tightly integrated parametric characterization system, including the measuring instruments and the test fixture, probe station, switching system, connections, cabling, grounding, and shielding. Within this “system view,” several factors can significantly affect a system’s overall performance. Even with a properly configured system, measurement noise and accuracy can be affected by:

- Grounding, shielding, and guarding,
- Connections at a thermal chuck,
- Instrument noise and settling time,
- Offsets in switch matrix contacts,
- Cabling, and
- Probe card design.

Effective techniques for minimizing noise and other errors from these sources must be used when making ultra-low current measurements.

System grounding and shielding

A typical semiconductor characterization system includes DC source-measure units (SMUs), a switch matrix, a probe station, and cabling between various components. Core instrumentation is housed in a main chassis that contains a PC host controller, low noise power supplies, SMUs, related instrumentation boards, and an assortment of data communication ports.

With this instrumentation arrangement, it is important to distinguish between the different grounds available. The main system chassis typically has an instrument common and a chassis power ground. The instrument common is the ground for the complete measurement circuit; it affects the system's low level measurement performance. In contrast, the chassis ground is connected to the power line ground and is mainly used for safety.

When the system is shipped from the factory, these two distinctly different grounds usually are connected. While this generally does not create a problem, there is the potential for noise errors. Sometimes, the power line ground can be noisy. In other cases, a test fixture and probe station connected to the instrument may create a ground loop that generates additional measurement noise. Because of these potential problems, low level measurement accuracy demands that system ground connections be thought out carefully.

Technically, although grounding and shielding are closely related, they are two different issues. In a test fixture or probe station, the device under test (DUT) and probe are typically enclosed in soft metal shielding. This metal enclosure eliminates interference from power lines and high frequency radiation (i.e., RF or microwave) and reduces magnetic interference. Most semiconductor devices are light-sensitive, so the enclosure also shades the DUT to prevent light-induced current flow that would interfere with accurate low level current measurements.

The metal enclosure is normally grounded for safety reasons. When an instrument is connected to the probe station through triaxial cables, however, the proper ground connection point is very important. A common grounding design error is to connect the instrument common and the chassis ground (**Figure 1a**). The illustration also shows the probe station grounded to the power line locally. Even more significantly, the measurement instrument and the probe station are connected to different power outlets. The power line grounds of these two outlets may not be at the same potential at all times. This creates a ground loop in which a fluctuating current may flow between the instrument and the probe station. To avoid ground loops, a better grounding scheme uses a single point ground (**Figure 1b**).

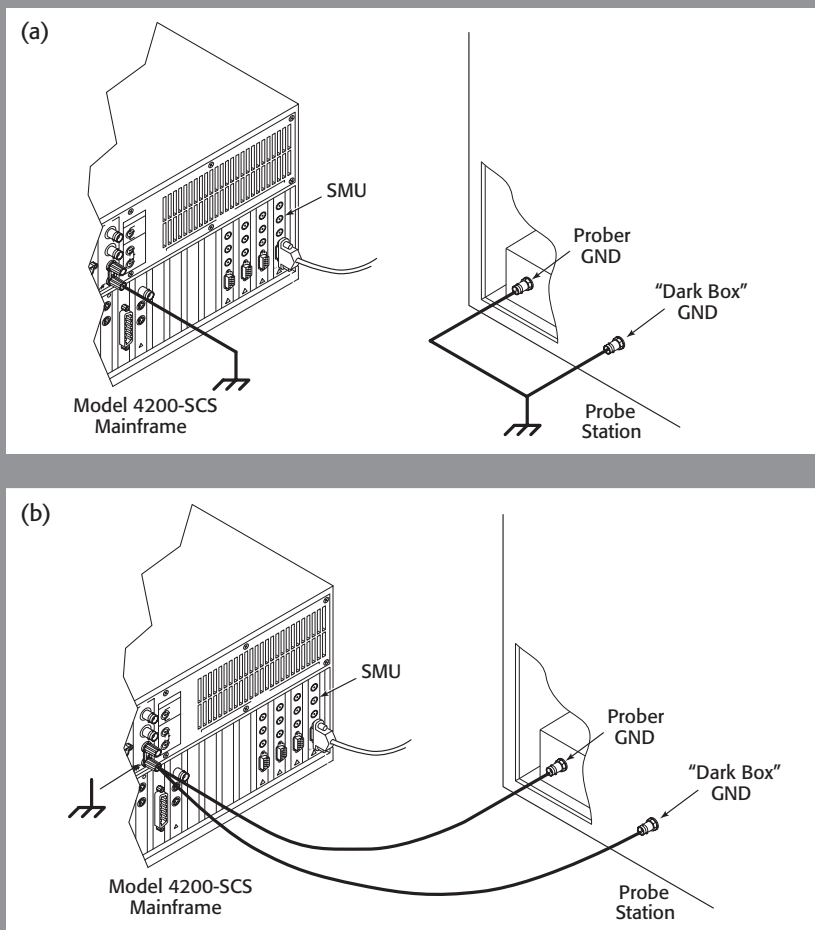


Figure 1. Problematic instrument ground loops (a) and connections that eliminate ground loops (b).

Minimizing system noise

Even when a characterization system is properly shielded and grounded, it is still possible for noise to appear in measurement results. Typically, these systems contribute very little noise to overall error. For example, the noise specification for Keithley's Model 4200-SCS semiconductor characterization system is only $\sim 0.2\%$ of range (i.e., peak-to-peak noise on the lowest range is just a few femtoamps). Noise can be further

reduced with proper signal averaging (e.g., through filtering or increasing power line cycle integrations).

The more likely sources of noise are other components of the system, such as long cables or switching hardware that are inappropriate for the application. Thus, it is advisable to use the best switch matrix available, designed specifically for ultra-low current measurements. Then, keep all connecting cables as short as possible.

Generally, system noise has the greatest impact on measurement integrity when the DUT signal is very small (i.e., low signal-to-noise ratio). This leads to the classic problem of amplifying noise along with the signal. Clearly, the key to low level measurement accuracy is to increase the signal-to-noise ratio.

Some characterization systems offer a low noise pre-amplifier option that enables sub-femtoamp measurement. To get that level of sensitivity, it is best to mount pre-amps remotely on a probe station platen. With this arrangement, the signal must travel only a very short distance (just the length of the probe needle plus a very short length of interconnect cabling, typically 6cm) before it is amplified. Then, the amplified signal is routed through the cables and switch matrix into the measurement hardware.

Both ambient temperatures and thermal chucks are used in device characterization measurements. Each approach offers a choice of regular (coaxial) or guarded (triaxial) chuck designs. Coaxial chucks generally are less costly, but coaxial connections tend to have higher leakage and noise than triaxial designs. Ultra-low current measurements require guarded chuck connections using triaxial cables.

In a guarded thermal chuck connection, the effects of leakage currents are reduced by forcing the inner shield (guard) of the triaxial cable to the same voltage as the center conductor of the cable. This is accomplished with a unit gain guard buffer in the SMU (**Figure 2**). The guard sense lead is used to detect the potential on the center conductor.

The heating element found in thermal chucks tends to add considerable noise to the measurement, especially if it uses AC current. Whenever possible, a DC heater should be used to minimize noise introduced into the measurement. To reduce noise still further, the heater can be switched off during the measurement once the desired temperature has been reached. The drawback to this technique is a slight drop in temperature after the heater is turned off. Still, if the measurement can be performed quickly, accuracy should not suffer significantly.

Measurement settling errors

Overall accuracy of the characterization system is affected by the way in which system elements work together. A key part of this integration issue is synchronizing measure-

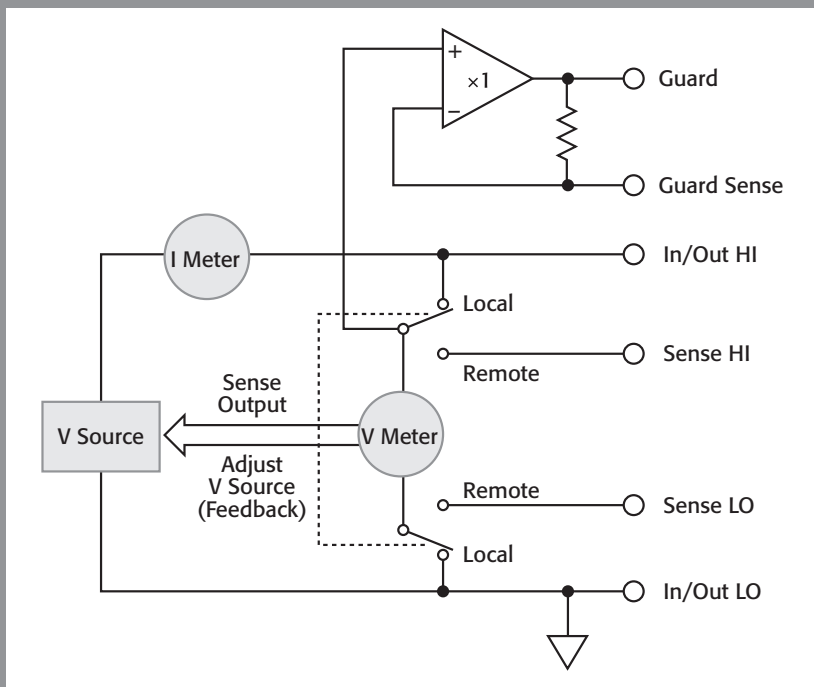
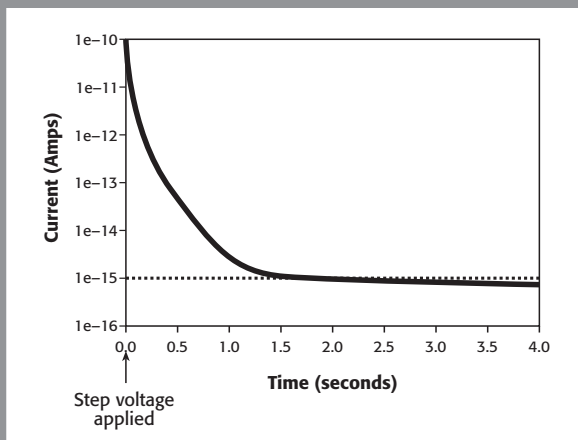


Figure 2. Simplified block diagram of an SMU configured to source voltage and measure current.

ment instruments with the switch hardware and probe. This is crucial in high speed or low current measurements. With improper synchronization and source-measure delay time, an erroneous signal such as measurement settling current may be collected, instead of the real device parameter. Signal averaging cannot eliminate errors introduced by settling currents.

A step voltage test is typically used to characterize system settling time. A 10V step is applied across two open-circuit probe tips, then current is monitored continuously for a period of time. The resulting current vs. time (I-t) curve (**Figure 3**) reveals several important system characteristics. From the I-t curve, it is possible to determine the amount of time required to reach a certain current level. For instance, if a 100fA current level were required by a particular test, the amount of settling time needed to reach this level could be determined simply by finding the 100fA level and locating the corresponding elapsed time on the x axis. Similarly, when testing several devices on a wafer, the same

Figure 3. Use SMU settling time to set source-measure delays. Leakage current reveals the limit on basic instrument sensitivity (here, 10^{-15}A).



technique can be applied to characterize the step response of a DUT, and the resultant I-t curve can be used to tune the delays, making it possible to reduce the test time on subsequent devices, as well as improve the accuracy of the measurement.

There are two regions of interest on the I-t curve: the transient segment and the steady-current segment. Immediately after the voltage step, the transient current will gradually decay to a steady value. The time it takes the current to reach that value is system settling time. Typically, the amount of time needed to reach $1/e$ of the initial value is defined as the system time constant. The time constant can vary widely for different systems, equipment, and cabling configurations.

Settling time is mainly the result of capacitance inherent in switch relays, cables, etc. It may also be a function of dielectric absorption that occurs in the insulating materials of various system components. If system components use materials with high dielectric absorption, then the system settling time may be quite long.

The longer the settling time, the longer it takes to get an accurate measurement. Conversely, an extended settling time can lead to inaccurate readings in high speed measurements. For example, if the source-measure delay is very short during a leakage current test on high resistivity material, a significant portion of the measurement is actually transient current, not the leakage current. This may lead to the erroneous conclusion that the material is excessively leaky. In addition to this error, another potential problem is wide fluctuations that sometimes occur in transient currents. These fluctuations may be misinterpreted as the result of a noisy system.

Once the transient current has settled to its steady value, it corresponds to the system leakage current. Typically, system leakage current is expressed as amperage/volt. To determine its magnitude, simply measure the steady-state current and divide by the voltage step value.

Since most low level characterization systems now use low noise triaxial cables, leakage current is rarely the result of cabling. Leakage current typically comes from switch relays or the probe card. Two types of leakage are associated with the switch and probe card:

- Path-to-ground leakage occurs in the path from the relay to the instrument ground or from the probe card pin to the ground.
- Path-to-path is the leakage between adjacent switch relays or probe card pins.

The magnitude of the system leakage current establishes the noise floor and overall sensitivity of the system.

Proper measurement speed

For the best possible throughput, it is desirable to shorten the SMU source-measure delay time to the minimum required for acceptable accuracy. This is a function of not only system settling time, but also SMU characteristics. Unless the system settling time is much shorter than SMU latencies — which is unlikely — the system settling time leakage-current I-t curve will be the primary tool for determining proper source-measure delay.

With the system leakage I-t curve in hand, the next step is to establish acceptable measurement sensitivity. Suppose the task requires accurate DUT leakage measurements only at the picoamp level. Then, source-measure delay time can be established by a point on the transient portion of the system settling curve where the leakage current is at a sub-picoamp level. If the expected DUT current is in the tens of femtoamps range, then the delay time must be lengthened, so the transient current reaches a lower value before the measurement is taken.

Appropriate system elements

A parametric characterization system can only be as good as its worst-performing component. For instance, if a good instrument is integrated with a poor performing switch or prober, the latter determines overall measurement performance, no matter how good the instrument. The following examples illustrate the importance of proper component selection in ensuring system accuracy.

A Keithley Model 4200-SCS has 100fA current resolution and 10fA current measurement accuracy, but if the system is configured with a switch card that has 100fA offset

current (such as the Keithley Model 7072A semiconductor matrix card), its measurement accuracy will be limited to approximately 100fA, and extended settling times will be required. Newer “air matrix” technology used in Keithley’s Model 7174A card provides significantly better leakage performance (10fA) and faster settling times (typically, <2.5 seconds to 400fA after a 10V step).

To maintain low settling times, short cables should generally be used in the system. Excessive cable lengths slow down measurements and test system throughput. In addition, all cables are not created equal, even when they carry a triaxial designation. When in doubt, order cables that have been tested and matched to the instruments being used, which usually are available from the instrument vendor.

The performance of probe cards and manipulators also varies considerably. While epoxy-based cards are suitable for many applications and are relatively economical, they have higher leakage current and longer settling time specifications than newer card designs. For ultra-low current measurements, two-layer probe cards with Teflon®-insulated coaxial feed-throughs and ceramic blade needle mounts provide minimal leakage and dielectric absorption.

If a manipulator is needed, consider using a specially designed triaxial guarded manipulator. While a coaxial manipulator may be more economical or settle slightly faster, the low current performance of this design typically doesn’t match that of a triaxial manipulator.

Integration issues

Better accuracy and throughput can be achieved with careful selection of parametric characterization system elements. Then, appropriate grounding and shielding techniques for chucks, switches, cabling, probe cards, etc. will take advantage of the accuracy and speed inherent in these instruments.

Still, ease of integration and usage can have a strong effect on productivity. For example, look for easy integration of user-written C language subroutines into the test environment. These routines can control the internal SMUs and external instruments, such as pulse generators and C-V meters, connected on a general purpose interface bus or RS-232 data bus. If a pre-amp is needed, its integration into the system should simply appear to a user as if the SMU has additional ranges. Of course, the system software must support switch matrixes and analytical probes (both automatic and manual types). Integration of all these system features and instruments should allow control of the entire test rack from a single user interface.

The overall approach

By taking an integrated approach to making ultra-low current measurements with parametric test hardware, overall efficiency of the test process can be achieved. One can

achieve better low current measurements by analyzing the system from instruments to the DUT and identifying the potential sources of error. Once the sources of error are known, the solution can be tailored to the budget.

For example, fixing the current range can in many cases reduce the testing time and requires no change in hardware. Sometimes, a direct connection to the DUT that bypasses the switch matrix can improve performance. If this is not practical, choosing a specialized low current switch card will help. Achieving the ultimate in low current performance requires a top-down, system-oriented approach.

Upgrading parametric test system performance generally involves a one-time expense. Some of our recommendations cost little or nothing, depending on test hardware features. Guarding techniques may require a shift from coaxial to triaxial cables, which is a modest expenditure. Determining the system settling time to optimize accuracy and throughput requires a few hours of testing. A low noise SMU pre-amp will increase system sensitivity for ~\$2000.

If system performance is still less than desired after these modifications, then replacement of existing switch matrix cards with low leakage types should be considered; this can cost up to a few thousand dollars for a low current, high speed 8×12 matrix card. Replacement of a coaxial AC thermal chuck with a triaxial DC model could cost several thousand dollars. Switching from a coaxial manipulator to a triaxial guarded design could run as high as \$2000.

In many cases, it may be less costly to rebuild or upgrade existing system components such as a chuck or manipulator to a higher performance level, instead of buying new equipment. While each test situation is different, these optimization strategies can significantly improve test throughput and accuracy, lower the cost of testing, and provide an excellent return on investment. ■

OVERCOMING THE MEASUREMENT CHALLENGES
OF ADVANCED SEMICONDUCTOR TECHNOLOGIES

SECTION VI

**Appendix A:
Selector Guides**

Semiconductor Test Solutions

From materials research to production test and monitoring, Keithley is the leader in semiconductor test solutions that IC makers trust. In the development lab, Keithley semiconductor characterization systems provide unmatched sensitivity and flexibility for investigating material properties, describing device attributes, and qualifying new designs. Keithley is also at the leading edge of automated parametric test technology for production process monitoring and has pioneered many innovations that shorten test cycles and lower the cost of test.

Materials Research Solutions

Our innovative instruments and systems are helping researchers commercialize technologies like high κ gate materials, copper traces, low κ insulation, III-V compounds, and carbon nanotubes. Their modular design allows greater integration flexibility and simplifies expanding and repurposing systems to address new test needs cost-effectively.

Material	Application Needs	Keithley Solution
Conductor/ interconnect materials	Low resistance and capacitance measurements	Model 2182A Nanovoltmeter Model 2010 DMM Model 2750 Multimeter/Switch System Model 4200-SCS Series 2400/6430 SourceMeter Instruments Model 590 C-V Analyzer
	Low current measurements	Model 6485 Picoammeter Model 4200-SCS Model 6430 Sub-Femtoamp Remote SourceMeter Instrument
Superconductors	Fast, low noise measurements of low voltages Fast, low noise measurements of low resistances	Model 2182A Nanovoltmeter Model 1801 Preamp Series 2000 DMMs Model 2510 TEC SourceMeter Instrument
Insulators/ dielectrics	High resistance measurements	Model 6517A Electrometer Model 6430 Sub-Femtoamp Remote SourceMeter Instrument Model 6485 Picoammeter
Semiconductors	Stress test for reliability Capacitance (charge storage) Resistivity measurements	Model 4200-SCS Model 6430 Sub-Femtoamp Remote SourceMeter Instrument Model 590 C-V Analyzer Model 707A/7174A Switch System Series 2000 DMMs S510 Semiconductor Reliability Test System

Device Characterization Solutions

Semiconductor manufacturers rely on Keithley instruments and systems to help them get to market faster, hit targeted yields quicker, and achieve the highest device reliability in the shortest possible time. Systems are designed specifically for critical tests that semiconductor fabs and researchers need most.

Device Under Test	Application Needs	Keithley's Solution
Transistors	I-V, C-V measurements	Model 4200-SCS Model 590 C-V Analyzer Series 2400 SourceMeter Instruments
Device Modeling	Data acquisition and parameter extraction	Model 4200-SCS Model 590 C-V Analyzer
Capacitors	Capacitance, charge, low current, breakdown voltage	Model 4200-SCS Model 2410 SourceMeter Instrument Model 590 C-V Analyzer Model 6517A Electrometer
Nanoelectronic Devices	C-V characterization Low resistance measurements Nanoelectronic and molecular electronic I-V curves	Model 4200-SCS Model 6517A Electrometer Model 2182A Nanovoltmeter Model 590 C-V Analyzer Model 6430 Sub-Femtoamp Remote SourceMeter Instrument
Diodes	High breakdown voltage, I-V curves	Model 4200-SCS Models 2400/2410 SourceMeter Instruments
Resistors	Wide dynamic range, high testing accuracy, fast measurements	Model 4200-SCS Series 2400 SourceMeter Instruments Model 6517A Electrometer Series 2000 DMMs

Reliability Test Solutions

Keithley supplies integrated solutions for reliability testing that help manufacturers control yields, reduce field failures, and increase customer satisfaction. From wafer-level devices to packaged parts testing, these systems come with superior measurement algorithms for exceptional throughput.

Test	Application Needs	Keithley's Solution
Accelerated Stress Testing (AST)	Source power, monitor temperature, measure multiple DUTs	Model 2400 SourceMeter Instrument Integra Series Systems
Quality Assurance Testing (QAT)	Modular building blocks for custom-built solutions and drivers	Model 2400 SourceMeter Instrument Model 6430 Sub-Femtoamp Remote SourceMeter Instrument Integra Series Systems Series 700 Switch Systems Series 2000 DMMs LabVIEW VI Drivers, VISA Drivers
Monitor and log lab environmental parameters	Temperature and humidity measurement, distributed/remote communication	Integra Series Systems
Wafer Level Reliability	Stress multiple transistors in parallel and measure small degradations	Model 4200-SCS with KTEI 5.0 or later S510 Semiconductor Reliability Test System

Process Monitoring Solutions

Keithley helps fabs control their parametric test costs without compromising throughput or accuracy. A variety of system features make this possible, including ultra-low leakage signal paths, parallel (single insertion) DC and RF measurements, adaptive testing with flexible sampling, and Recipe Manager software that speeds generation of valid test plans.

Process	Application Needs	Keithley's Solution
New technologies	RF/DC, SOC, FeRAMs, MRAMs, LCD-TEGs	S400/S600 Series Parametric Test Systems Model 4200-SCS
Gate/poly	MOSCAP GOI, ECD, V_t	Model 4200-SCS S400/S600 Series Parametric Test Systems Model 2400 SourceMeter Instrument
Metal-2	Contact check, electromigration, EWR	S400/S600 Series Parametric Test Systems Model 4200-SCS
Equipment qualification	Defect density	S400/S600 Series Parametric Test Systems
End-of-line wafer acceptance	Transistors (V_t), diodes, capacitors, resistors, gate delay	S400/S600 Series Parametric Test Systems Model 4200-SCS

Functional Test Solutions

Economical, high speed functional testing is available from Keithley, even for low volume applications. Highly repeatable pass/fail testing of packaged devices, like RFICs, optical ICs, and other short-run or specialty components, helps reduce manufacturers' total Cost of Test. Expertise in system integration allows Keithley to incorporate parts handling and high speed switching for a variety of applications.

Process	Application Needs	Keithley's Solution
IDDQ (SOC and other highly integrated devices)	Source V, measure low I	Model 2400, 2420 SourceMeter Series 7000 Switch Mainframes Series 7000 Switch Cards Model 4200-SCS Model 4500-MTS Series 2600 System SourceMeter Instruments
Optical ICs and transceivers	Source V, measure I Source I, measure V	Model 2425 100W SourceMeter Model 2520 Pulsed Laser Diode Test System Series 7000 Switch Mainframes Series 7000 Switch Cards Model 4500-MTS Series 2600 System SourceMeter Instruments
Customized flexible low volume functional test	Test sequence list, fast pass/ fail, binning/sorting, RF	Series 2400 SourceMeter Instruments Model 707A/708A Switch Mainframes Series 7000 Switch Cards System 40 RF/Microwave Signal Routing Systems Model 4500-MTS Series 2600 System SourceMeter Instruments
Specialized testing	Multi-channel I-V testing	Model 4500-MTS Series 2600 System SourceMeter Instruments

Source and Measure Products Selector Guide

MODEL	2400 2400-C 2400-LV	2410 2410-C	2420 2420-C	2425 2425-C	2430 2430-C	2440 2440-C
Description	General Purpose	High Voltage	3 A	High Power	Pulse	5 A
Current Source/Sink	•	•	•	•	•	•
Voltage Source/Sink	•	•	•	•	•	•
POWER OUTPUT						
	22 W	22 W	66 W	110W	1100 W*	55 W
CURRENT CAPABILITY						
Min.	±10 pA	±10 pA	±100 pA	±100 pA	±100 pA	±100 pA
Max	±1.05 A	±1.05 A	±3.15 A	±3.15 A	±10.5 A*	±5.25 A
VOLTAGE CAPABILITY						
Min.	±1 µV	±1 µV	±1 µV	±1 µV	±1 µV	±1 µV
Max.	±21/±210 V	±1100 V	±63 V	±105 V	±105 V	±42 V
OHMS RANGE						
	<0.2 Ω to >200 MΩ	<0.2 Ω to >200 MΩ	<0.2 Ω to >200 MΩ	<0.2 Ω to >200 MΩ	<0.2 Ω to >200 MΩ	<2.0 Ω to >200 MΩ
BASIC ACCURACY						
I	0.035%	0.035%	0.035%	0.035%	0.035%	0.035%
V	0.015%	0.015%	0.015%	0.015%	0.015%	0.015%
Ω	0.06 %	0.07 %	0.06 %	0.06 %	0.06 %	0.06 %
FEATURE SUMMARY						
Programming	IEEE-488, RS-232					
Memory	5000 point, 2500 point reading buffer					
Trigger	Trigger Link with 6 In/Out					
Digital I/O	1 In/4 Out with built-in component handler interfaces.					
Guard	Ohms (high current) and cable					
Other	5½-digit measure capability. Handler interface. 500µs pass/fail test. Optional contact check capability.					
Compliance	CE and UL	CE	CE	CE	CE	CE

* In pulse mode. ** 1aA = 1×10^{-18} A. *** Approximate average.

APPLICATIONS

2400: Resistive devices, diodes, optoelectronic components, IDDQ testing.

2410: Voltage coefficient, varistors, high voltage diodes and protection devices, airbag inflators.

2420: Power resistors, thermistors, solar cells, batteries, diodes, IDDQ testing.

2425: Power semiconductors, DC/DC converters, high power components, IDDQ testing.

2430: High power pulse testing; varistors and other circuit protection devices.

2440: 5A pump laser diodes.

6430	2601 2602	4200-SCS	4500-MTS	S510	
Ultra-Low Current	Single/Dual Channel System SourceMeter Instruments	Multi-Channel I-V Characterization	Multi-Channel I-V Production Test	Semiconductor Reliability Test System	
				High Resolution	High Speed Parallel
•	•	•	•	•	
•	•	•	•	•	•
POWER OUTPUT					
2 W	40.8 W/ch.	Up to 96.8 W	Up to 216 W	2 W/ch.	100mW/ch.
CURRENT CAPABILITY					
±10 aA**	±1 pA	±100 aA w/ preamp	±0.1 nA	±1 fA	±100 pA
±105 mA	±3 A	±1 A w/4210-SMU	±1 A	±100 mA	±10 mA
VOLTAGE CAPABILITY					
±1 µV	±1 µV	±1 µV	±10 mV	±1 µV	±10 µV
±210 V	±40.4 V/ch.	±210 V	±10 V	±210 V	±10 V
OHMS RANGE					
<2.0 Ω to >20 TΩ	N/A	N/A	N/A	N/A	
BASIC ACCURACY					
0.035%	0.02%	0.05 %***	0.065%	0.15%*	0.3%
0.015%	0.015%	0.012%***	0.06 %	0.05%*	0.3%
0.06 %	N/A	N/A	N/A	N/A	
FEATURE SUMMARY					
IEEE-488, RS-232	IEEE-488, RS-232	Embedded GUI	Ethernet	Embedded GUI	
5000 pt, 2500 pt reading buffer	200,000 readings per SMU	4096 sample memory per card	Up to 1,000,000 points per card	4096/ch.	125,000/ch.
Trigger Link with 6 In/Out	14 general I/O trigger lines	Internal only	Triggering w/PCI digital I/O	N/A	
Same as 24xx		N/A	With opt. PCI card	N/A	
Same as 24xx	Cable	Cable	Cable	Cable	
Same as 24xx	Scalable from 1 to 16 channels with TSP-Link®	Optimized for front panel operation	5½-digit measure. Handler interface w/PCI digital I/O.	Optimized for ULSI Si CMOS Gate Stack Reliability Testing	
Yes	CE and UL	Yes	Yes	Yes	

APPLICATIONS

6430: Particle beam experiments, SET (single electron transistor) testing, ultra-high resistance testing (up to $10^{15}\Omega$).

2601, 2602: Characterization and functional test of a wide range of semiconductor components, including discretes, optoelectronics, integrated circuits, and RFICs.

4200-SCS: Device characterization, parametric I-V analysis, stress-measure, reliability testing, device modeling, materials research.

4500-MTS: High speed, parallel testing of HiB LEDs, resistive MEMs, photonic integrated circuits, optoelectronic devices, packaged parts, radio frequency integrated circuits (RFICs).

S510: Automated WLR test.

Low Current/High Resistance Measurements Selector Guide

	Current Amplifier	Picoammeters		
MODEL	428	6485	6487	2502
CURRENT MEASURE				
From ¹	1.2 fA	20 fA	20 fA	15 fA
To	10 mA	20 mA	20 mA	20 mA
VOLTAGE MEASURE				
From ²				
To				
RESISTANCE MEASURE⁴				
From ⁵			10 Ω	
To ⁶			1 P Ω	
CHARGE MEASURE				
From ²				
To				
FEATURES				
Input Connection	BNC	BNC	3 Slot Triax	3 Slot Triax
IEEE-488	•	•	•	•
RS-232		•	•	•
Guard				
CE	•	•	•	•
Other	2 μ s rise time. 10 ¹¹ V/A gain.	5½ digits. Autoranging. 1000 rdg/s.	5½ digits. Built-in 500V source. Alternating voltage method for HI-R sweeps.	5½ digits. Dual channel. Built-in 100V source per channel.

1. Includes noise.
2. Digital resolution limit. Noise may have to be added.
3. P Ω (Petaohms) = 10¹⁵ Ω .
4. Resistance is measured with the 236, 237, and 238 using Source V/Measure I or Source I/Measure V, but not directly displayed.
5. Lowest resistance measurable with better than 1% accuracy.
6. Highest resistance measurable with better than 10% accuracy.

Electrometers			Source-Measure Units		
6514	6517A	6430	236	237	238
CURRENT MEASURE					
<1 fA	<1 fA	400 aA	30 fA	30 fA	30 fA
20 mA	20 mA	100 mA	100 mA	100 mA	1 A
VOLTAGE MEASURE					
10 μ V	10 μ V	10 μ V	10 μ V	10 μ V	10 μ V
200 V	200 V	200 V	110 V	1100 V	110 V
RESISTANCE MEASURE⁴					
10 Ω	100 Ω	100 $\mu\Omega$	100 $\mu\Omega$	100 $\mu\Omega$	50 $\mu\Omega$
200 G Ω	10 P Ω^3	10 P Ω^3	0.1 P Ω^3	1 P Ω^3	0.1 P Ω^3
CHARGE MEASURE					
10 fC	10 fC				
20 μ C	2 μ C				
FEATURES					
3 Slot Triax	3 Slot Triax	3 Slot Triax	3 Slot Triax	3 Slot Triax	3 Slot Triax
•	•	•	•	•	•
•	•	•			
•	•	•	•	•	•
•	•	•	•	•	•
5½ digits. Replaces Models 6512, 617-HIQ.	5½ digits. Built-in ± 1 kV source. Temperature, RH measurements. Alternating polarity method for HI-R. Plug-in switch cards available.	SourceMeter with Remote PreAmp to minimize cable noise.	Source/measure capability. High speed. 5 digits.		

Sourcing Solutions for Low Current/High Resistance Measurements

MODEL	6220	6221	248
Current Source	•	•	
Voltage Source			•
Sink	•	•	•
CURRENT OUTPUT			
Accuracy ¹	2 pA	2 pA DC, 4 pA AC	
Resolution ²	100 fA	100 fA (DC & AC)	
Maximum	±105 mA	±105 mA	
VOLTAGE OUTPUT			
From			±1.5 V
To			±5000 V
POWER OUTPUT			
	11 W	11 W	25 W
CURRENT LIMIT			
			5.25 mA
VOLTAGE LIMIT			
	105 V	105 V	0 to 5000 V
ACCURACY (±Setting)			
I	0.05%	0.05%	
V			0.01%
FEATURES			
Output Connector	3 Slot Triax	3 Slot Triax	SHV High Voltage Coax
Ethernet		•	
RS-232	•	•	
IEEE-488	•	•	•
Memory	65,000 pt.	65,000 pt.	
Remote Sense			
Current Source Guard	•	•	
CE	•	•	•
Other	Controls 2182A for low power resistance and I-V measurements.	AC and DC current source. ARB waveforms up to 100kHz. Controls 2182A like 6220, adds pulsed I-V.	Voltage monitor output. Programmable voltage limit.

1 Best absolute accuracy of source.

2 Resolution for lowest range, smallest change in current that source can provide.

236	237	238	4200-SCS	6430
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
CURRENT OUTPUT				
450 fA	450 fA	450 fA	40 fA	10 fA
100 fA	100 fA	100 fA	1.5 fA	50 aA
±100 mA	±100 mA	±1 A	1.05 A	±105 mA
VOLTAGE OUTPUT				
±100 µV	±100 µV	±100 µV	±5 µV	±5 µV
±110 V	±1100 V	±110 V	±210 V	±210 V
POWER OUTPUT				
11 W	11 W	15 W	22 W for 4210-SMU 2.2 W for 4200-SMU	2.2 W
CURRENT LIMIT				
1 pA to 100 mA	1 pA to 100 mA	1 pA to 1 A	100 fA to 1.05 A	1 fA to 105 mA
VOLTAGE LIMIT				
1 mV to 110 V	1 mV to 1100 V	1 mV to 110 V	20 mV to 210 V	0.2 mV to 210 V
ACCURACY (±Setting)				
0.05%	0.05%	0.05%	0.06%	0.03%
0.03%	0.03%	0.03%	0.02%	0.02%
FEATURES				
Two 3 slot triax	Two 3 slot triax	Two 3 slot triax	Dual triax for 4200-PA. Dual mini-triax for 4200-SMU and 4210-SMU.	3 slot Triax
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
1000 pt.	1000 pt.	1000 pt.	4096 pt. / unlimited	2500 pt.
•	•	•	•	•
•	•	•	•	•
•	•	•	•	•
Source/measure capability. Pulse mode. High speed. Built-in waveforms.			GUI interface. Real-time graphing.	

OVERCOMING THE MEASUREMENT CHALLENGES
OF ADVANCED SEMICONDUCTOR TECHNOLOGIES

SECTION VII

**Appendix B:
Glossary**

Acceptance Testing

Testing performed on a product or equipment to determine whether an individual lot of the product or equipment conforms to specified requirements.

AEM

Analytical electron microscopy.

Alternative Dielectrics

Dielectrics with κ (dielectric constant) > 3.9 (dielectric constant of SiO_2) and acting as gate oxides in silicon MOS devices instead of SiO_2 , which are often referred to as “high κ dielectrics.” Dielectrics with $\kappa < 3.9$ and used as ILD are referred to as “low κ dielectrics.”

American National Standards Institute (ANSI)

An organization that compiles and publishes computer industry standards.

Avalanche Breakdown

Junction breakdown due to current increase caused by avalanche multiplication of charge carriers in the region featuring very high electric field.

Avalanche Multiplication

Generation of electron-hole pairs due to impact ionization in a depleted region of a semiconductor (e.g., in a reverse-biased p-n junction), which causes other impact ionization events, further increasing the number of carriers.

Back End Of Line Processes (BEOL)

Operations performed on the semiconductor wafer in the course of device manufacturing following first metallization.

Bias Temperature Stress (BTS)

Reliability test in which device is electrically stressed at increased temperature for a sustained time. Commonly used in C-V characterization of MOS devices.

Bipolar Device

Semiconductor device in which operation is based on the use of both majority and minority charge carriers. All p-n junction based devices fall into this category.

Bipolar Junction Transistor (BJT)

A transistor consisting of three semiconductor regions (emitter, base, and collector) with alternating conductivity type (i.e., n-p-n or p-n-p). A current controlled device. A key transistor structure in semiconductor electronics.

Breakdown

Catastrophic effect occurring in the presence of high electric field and causing originally high resistance element (e.g., MOS capacitor of reverse-biased p-n junction) to allow flow of high current. This is typically an irreversible effect, permanently damaging the element. It also occurs in materials in the presence of very high electric field.

Breakdown Voltage

Voltage at which breakdown occurs and current increases uncontrollably (unless limited by the external circuit).

Burn-In

The process of exercising an integrated circuit at elevated voltage and temperature. This process accelerates failure normally seen as “infant mortality” in a chip.

Capacitance-Voltage (C-V) Measurements

Several parameters of semiconductor materials and structures can be determined by measuring C-V characteristics. These measurements are routinely used for process diagnostics and monitoring in MOS technology. They allow characterization primarily of dielectric layers such as interface trap density, fixed charge, and oxide charge.

Carrier Generation

The process of creating electron-hole pairs in a semiconductor.

Carrier Injection

The process of introducing charge carriers from one region within semiconductor device to another (e.g., in a forward-biased p-n junction, electrons are injected to the p-type region and holes are injected to the n-type region).

Charge-To-Breakdown (Q_{BD})

A measure of reliability of oxides in MOS gates.

Cost Of Ownership (COO)

The cost of implementing a given process technology, including cost of equipment, maintenance, materials, impact on existing tool set, etc. Typically determined for a specific tool or set of tools needed to carry out a specific processing goal.

Current-Voltage (I-V)**Measurements**

Used to determine electrical characteristics of semiconductor test structures and devices by measuring current flowing across the device as a function of applied voltage.

Damascene

A process in which interconnect metal lines are delineated in dielectrics, isolating them from each other by means of chemical-mechanical planarization (CMP). In this process, the interconnect pattern is first lithographically defined in the layer of dielectric, and then metal is deposited to fill the resulting trenches. Then, excess metal is removed by chemical-mechanical polishing/planarization.

Device Relaxation

Relaxation is the temporary recovery of the device performance after the voltage stress is removed. This phenomenon may be caused by a reversible electrochemical reaction. Sometimes called *device recovery*.

Dielectric

A solid displaying insulating properties (energy gap typically wider than about 5eV). Its uppermost energy

band is completely empty; therefore, dielectric features extremely low conductivity. Fundamental characteristics of a dielectric are independent of the applied voltage. The most commonly used dielectrics in semiconductor technology are SiO_2 and Si_3N_4 .

Dielectric Relaxation Time

The time needed by a semiconductor to return to electrical neutrality after carrier injection or extraction.

Dual Damascene

A modified version of the damascene process, which is used to form metal interconnect geometry using a CMP process instead of metal etching. In dual damascene, two interlayer dielectric patterning steps and one CMP step create a pattern that would require two patterning steps and two metal CMP steps when using a conventional damascene process.

Field to Breakdown (E_{BD})

Electric field in the MOS gate oxide at which it breaks down, i.e., loses its insulating properties irreversibly. E_{BD} is measured by ramping the voltage applied to MOS gate while monitoring the current flowing across the gate oxide. The voltage at which large current starts to flow is recorded (V_{BD}), and then divided by dielectric thickness to obtain E_{BD} .

Electromigration (EM)

The self-diffusion of metal along interconnects, caused by the current flow through the metal. Electromigration is caused primarily by frictional

force between metal ions and flowing electrons, which results in a break in the metal line. It is a common cause of malfunctions of aluminum or copper interconnect networks in integrated circuits.

Equivalent Oxide Thickness (EOT)

A number used to compare the performance of high κ dielectric MOS gates with the performance of SiO_2 based MOS gates; shows thickness of SiO_2 gate oxide needed to obtain the same gate capacitance as the one obtained with thicker than SiO_2 dielectric featuring higher dielectric constant κ ; e.g., EOT of 1nm would result from the use a 10nm thick dielectric featuring $\kappa = 39$ (κ of SiO_2 is 3.9)

Flatband Voltage (V_{FB})

Refers to MOS devices, the bias voltage at which there is no net electrical charge in the semiconductor and, therefore, no voltage drop across it; in band diagram, the energy bands of the semiconductor are horizontal (flat).

Front End Of Line Processes (FEOL)

Operations performed on the semiconductor wafer in the course of device manufacturing up to first metallization.

Gate

An electrode that regulates the flow of current in a metal oxide semiconductor transistor.

Gate Oxide

A thin, high quality silicon dioxide film that separates the gate electrode of a metal oxide semiconductor transistor from the electrically conducting channel in the silicon.

GOI

Gate Oxide Integrity. The term implies electrical “integrity” of gate oxide. Determined through various current/voltage/electric field stress and breakdown tests of MOS gate stacks.

Hafnium Oxide (HfO₂)

High κ dielectric considered for next generation MOS gates; $\kappa \sim 25$; limited thermal stability with silicon; thermally stable up to 700°C.

Hafnium Silicate (HfSiO₄)

High κ dielectric considered for next generation MOS gates, $\kappa \sim 15$ –18. Thermodynamically stable with silicon.

Hard Breakdown

An abrupt change in the measured gate current (I_g) during TDDDB testing. This change is at least 5–10 \times the previous measurement. Also called a *hard failure*.

High κ Dielectric

Dielectric material featuring dielectric constant (κ) higher than 3.9 (the κ of SiO₂). Used as gate dielectrics in MOS devices and in storage capacitors.

High Frequency C-V

A capacitance-voltage measurement

carried out at the frequency of 100kHz and higher (often 1MHz).

Hot Carriers

Carriers (either electrons or holes) that have been accelerated by the large traverse electric field between the source and the drain regions of a metal oxide semiconductor field-effect transistor (MOSFET). They can jeopardize the reliability of a semiconductor device when these carriers are scattered (deflected) by phonons, ionized donors or acceptors, or other carriers. The scattering phenomenon can manifest itself as substrate current, gate current, or trapped charges.

Landing Pad

Contact metal pads on a wafer usually associated with test probes and existing for test only.

Low κ Dielectric

Dielectric material featuring $\kappa < 3.9$ (the κ of SiO₂). Used to insulate adjacent metal lines (interlayer dielectric, ILD) in advanced integrated circuits. Low κ reduces undesired capacitive coupling, and hence “crosstalk” between lines.

Majority Carrier

A type of charge carrier constituting more than one-half the total charge carrier concentration (for example, holes in p-type material).

Minority Carrier

A type of charge carrier constituting less than one-half of the total charge-

carrier concentration (for example, electrons in p-type material).

Minority Carrier Lifetime

The average time interval between the generation and recombination of minority carriers of a homogeneous semiconductor.

Mobility

The ensemble average of drift velocity of a charge carrier per unit electric field of a carrier in a semiconductor. The unit customarily used is square centimeter per volt-second. The preferred SI unit is the square meter per volt-second.

Nanotube

Rollled sheet of hexagonal carbon structures in the “chicken wire”-like form. Can be either semiconducting or metallic. Due to their ultra-small size, they should, in theory, provide faster switching than any of today’s semiconductor structures.

Nanowire

A nanowire is a wire of dimensions of the order of a nanometer (10^{-9} meters). At these scales, quantum mechanical effects are important, so such wires are also known as “quantum wires.”

NBTI

Negative Bias Temperature Instability. NBTI is a phenomenon where change in the gate-channel interface causes degradation in pMOS device performance. During NBTI testing, the stress portion of the Stress/Measure

Cycle consists of applying a stress voltage ($V_{\text{stress}} > V_{\text{operation}}$) at an elevated temperature ($100^{\circ}\text{C}+$) to accelerate performance degradation to predict device lifetime. The degradation is typically tracked as the increase of the transistor voltage threshold (V_T), although other characteristics also degrade. This degradation can reduce yield through failures during burn-in or in the field. *Device relaxation* is a concern for NBTI.

Oxide

An insulating layer that is the product of an oxygen reaction with a given element. In silicon processing, this term is typically associated with SiO_2 .

Oxide Trapped Charge

Charge centers in SiO_2 and other gate dielectrics, which are electrically activated/de-activated by trapping/de-trapping charge carriers injected into the oxide either from the gate or from the substrate. Oxide trapped charge causes instabilities of MOSFET characteristics.

Process Control

The ability to maintain specifications of product and equipment during the manufacturing operations, typically done to maintain high product yield.

Quasistatic C-V

Capacitance-voltage measurement of MOS capacitor performed at very low frequency, e.g., $<50\text{Hz}$. When combined with high frequency (e.g., 1MHz) C-V measurements, it provides quantitative information regarding

the electronic properties of dielectric-semiconductor interface.

Radio Frequency (RF)

Electromagnetic energy with frequencies ranging from 3kHz to 300GHz. Microwaves are a portion of the radio frequency spectrum extending from 300MHz to 300GHz.

Relaxation

See *device relaxation*.

Scanning Electron Microscope (SEM)

Test equipment that uses an electron beam to display an electronically scanned image of a die or wafer for examination on a screen or for transfer to photographic film. Provides higher magnification than an optical microscope.

Shared Measure

Using switching to permit measuring more than one DUT signal. The measurements happen sequentially across the available DUTs, whereas there is typically a dedicated source(s) for each DUT. Contrast with shared SMU, where both the source and measure are switched.

Shared SMU

Using switching or DUT connection methods that permit more than one DUT to be connected to, or powered by, a single SMU.

SMU

Source-Measure Unit. An instrument or capability consisting of both sourcing a DC signal (such as current

or voltage) and measuring a signal (current or voltage, or both).

Soft Breakdown

Dramatic increase in the current noise in the measured gate current (I_g) during TDDDB testing. This change in noise may be accompanied by a gradual change in the magnitude. Also called a *soft failure*.

SOI

Silicon-On-Insulator. SOI may be the silicon substrate of choice for future generations of CMOS ICs. Basically, an SOI substrate is a silicon wafer with a thin layer of oxide (SiO_2) buried in it. Devices are built into a layer of silicon on top of the buried oxide. SOI substrates provide superior isolation between adjacent devices in an integrated circuit vs. devices built into bulk wafers. They also offer improved performance of SOI devices due to reduced parasitic capacitances.

STM

Scanning Tunneling Microscopy. This method allows visualizing conductive solid surfaces with atomic resolution. A conductive tip is scanned across the surface of the wafer. Information on the physical features of the surface is based on the measurement of the tunneling current between the tip and the surface atoms.

Stress/Measure Cycle

Reliability test method where voltage stress is applied for a specified period of time, after which a measurement is made. An entire test may consist

of from one to dozens of cycles. The stress interval periods may be equal in length or may increase logarithmically. Both NBTI and TDDB use the stress/measure cycle.

Stress-Switch-Measure

Method for sharing SMUs for reliability testing. The key is that all SMUs in the system are dedicated to either measurements or stressing. For FET DUTs, there are typically four SMUs for the FET (gate, drain, source, body) measurement and at least one for stressing.

Stress Monitor

Readings of gate and/or drain current during the stress portion of the stress/measure cycle. Applies to both NBTI and TDDB.

TDDB

Time Dependent Dielectric Breakdown. TDDB tests the lifetime of the gate dielectric via voltage stressing ($V_{\text{stress}} > V_{\text{operation}}$) over time at an elevated temperature. Unlike NBTI, where the phenomenon is looking for performance degradation, TDDB includes degradation as well as complete failure of the gate. TDDB consists of a single stress cycle, with gate current measurements tracking the degradation.

TEM

Transmission Electron Microscopy. Commonly used to visualize cross-sections of multi-layer nanostructures.

Trapped Charges

Charges trapped either in the gate oxide or, in the case of a lightly doped drain (LDD) metal-oxide semiconductor field-effect transistor (MOSFET), in the spacer region. Trapped charges in the gate or the spacer lead to threshold voltage shift or to transconductance degradation, respectively.

Wafer Level Reliability

A methodology for assessing the reliability impact of materials, tools and processes by testing mechanism-specific test structures under accelerated conditions during device processing.

Yield

The percentage of wafers, dice, or packaged units that conform to specifications. The most common yields in the manufacturing process are wafer fab yield (percentage of wafers that complete wafer processing), wafer probe yield (the fraction of dice on a wafer that meet the device specifications), assembly yield, and final test yield.



Want to learn more about Keithley's semiconductor parametric test and device characterization solutions?

Visit www.keithley.com/products/semiconductor to stay current on testing innovations from Keithley Instruments. Or, to discuss your application needs with a Keithley expert, contact one of the Keithley sales offices listed here.

KEITHLEY

Keithley Instruments, Inc.

Corporate Headquarters • 28775 Aurora Road • Cleveland, Ohio 44139
440-248-0400 • Fax: 440-248-6168 • 1-888-KEITHLEY (534-8453) • www.keithley.com

Belgium: Sint-Pieters-Leeuw • 02-363 00 40 • Fax: 02-363 00 64 • www.keithley.nl

China: Beijing • 8610-55010010 • Fax: 8610-82255018 • www.keithley.com.cn

Finland: Helsinki • 09-5306-6560 • Fax: 09-5306-6565 • www.keithley.com

France: Saint-Aubin • 01-64 53 20 20 • Fax: 01-60 11 77 26 • www.keithley.fr

Germany: Germering • 089/84 93 07-40 • Fax: 089/84 93 07-34 • www.keithley.de

Great Britain: Theale • 0118 929 7500 • Fax: 0118 929 7519 • www.keithley.co.uk

India: Bangalore: 91-80 2212 8027 • Fax: 91-80 2212 8005 • www.keithley.com

Italy: Milano • 02-48 39 16 01 • Fax: 02-48 30 22 74 • www.keithley.it

Japan: Tokyo • 81-3-5733-7555 • Fax: 81-3-5733-7556 • www.keithley.jp


Korea: Seoul • 82-2-574-7778 • Fax: 82-2-574-7838 • www.keithley.co.kr

Netherlands: Gorinchem • 0183-635333 • Fax: 0183-630821 • www.keithley.nl

Singapore: Singapore • 65-6747-9077 • Fax: 65-6747-2991 • www.keithley.com

Sweden: Solna • 08-509 04 600 • Fax: 08-655 26 10 • www.keithley.com

Taiwan: Hsinchu • 886-3-572-9077 • Fax: 886-3-572-9031 • www.keithley.com.tw



Specifications are subject to change without notice.
All Keithley trademarks and trade names are the property of Keithley Instruments, Inc.
All other trademarks and trade names are the property of their respective companies.

KEITHLEY

Keithley Instruments, Inc.

Corporate Headquarters • 28775 Aurora Road • Cleveland, Ohio 44139 • 440-248-0400 • Fax: 440-248-6168 • www.keithley.com

© Copyright 2005 Keithley Instruments, Inc.
Printed in U.S.A.

No. 2628
60515KQW