A Test-Structure to Efficiently Study Threshold-Voltage Variation in Large MOSFET Arrays

Nigel Drego, Anantha Chandrakasan, and Duane Boning
Microsystems Technology Laboratories, MIT, Cambridge, MA
{ndrego, anantha, boning}@mtl.mit.edu

Abstract

A test-structure comprising a dual-slope integrating analog-to-digital converter, auto-zeroing circuitry, digital control logic and a large array of Devices Under Test (DUTs) has been developed to isolate threshold voltage variation.. Threshold-voltage (V_T) isolation is achieved by testing all DUTs in the subthreshold regime where drain-to-source current is an exponential function of V_T. Spice simulations show that the structure is at least an order of magnitude more sensitive to V_T variation than to channel length variation. This, in combination with a hierarchical access scheme and leakage control system, allows efficient characterization of ΔV_T for ~70,000 NMOS and ~70,000 PMOS devices in a dense 2mm x 2mm DUT array.

1. Introduction

Process variation is increasingly becoming a limiting or determining factor in both IC design and manufacture [1]. Nearly all steps within the IC manufacturing process introduce varying degrees of variation in the end device due to limited controllability of the process module. In Deep-Sub-Micron (DSM) CMOS a number of steps can be highlighted as major sources of both random and systematic variation [2]: 1) sub-wavelength lithography, 2) plasma etch, 3) ion implantation and annealing, and 4) chemical-mechanical polishing (CMP) which is increasingly used for STI and poly planarity, and can result in pattern-dependent variation [3].

As gate dimensions continue to decrease, variation in ΔL and ΔV_T become increasingly worrisome due to decreasing depth-of-focus of sub-wavelength lithography, line-edge roughness, random discrete dopant fluctuation, stress effects, and oxide thickness (Δt_{ox}) fluctuation. Immersion lithography, extreme ultra-violet (EUV) lithography and improved resist materials may aid in improved control of ΔL. Solutions for improved control of random discrete dopant fluctuation or oxide thickness at the manufacturing level are problematic, meaning these are issues circuit designers and system architects must increasingly be aware of.

Recently, there has been an increasing drive to characterize, analyze and better understand the sources of variation as well as their circuit implications. A number of groups, including our own, have studied ring-oscillator frequency to characterize variation at both within-die and die-to-die levels [4][5][6]. While ring-oscillator based techniques enable ease of measurement, isolation of individual parameters for variability study is challenging due to amalgamation of the variation of many transistors into a single parameter (i.e. the frequency of ring operation). The authors of [6] are able to isolate V_T by including transistors in pass-gate configuration between each inverter stage of the ring. By using short rings, they are also able to limit the averaging occurring due to parameter lumping. Limiting placement of these rings to scribe-lines, however, does not enable the study of a large number of devices or within-die spatial correlation.

The work described in [9] also studies MOSFET threshold voltage mismatch in the sub-threshold domain but is limited to only 400 NMOS "cells" with off-chip current measurements. Furthermore, only the sigma in variation can be calculated and not the actual threshold voltage of each device, making within-die or spatial correlation analysis difficult.

In [7], an on-chip ammeter is built to enable collection of transistor I-V curves with digital I/O, enabling measurement of I-V characteristics of a larger number of devices than is typically sustained by common DC probing measurement schemes which require four pads per device. This design allows for complete simulation models to be created but requires tens of supporting transistors for each transistor being measured, resulting in fewer testable devices per unit area. The test structure in [11] also contains a large DUT-array for characterizing device mismatch but requires off-chip measurement and characterization.

While the design presented in this work possesses an architecturally similar approach to that in [7], there are fundamental differences: 1) isolation of V_T for study, 2) measurements solely within the sub-threshold regime, 3) source and sink DACs and associated logic to eliminate the effect of parasitic leakage currents from DUTs not being accessed, and 4) a much larger number of DUTs. Section 2 describes the high-level architecture and circuits necessary to extract threshold-voltage variation. Section 3 provides the analytic basis used to isolate ΔV_T for analysis and characterization, and Section 4 shows Spice simulation results of these circuits.

Proceedings of the 8th International Symposium on Quality Electronic Design (ISQED'07)
0-7695-2795-7/07 $20.00 © 2007 IEEE

Concluding remarks and future work are provided in Section 5.

2. Test-Structure Architecture

We first present a high-level array test-structure capable of accurately measuring sub-threshold currents of ~70K transistors of each type (PMOS and NMOS). The test structure allows for control of both V_{GS} and V_{DS} to enable current versus voltage measurements, while eliminating the effect of parasitic leakage currents from DUTs not being tested.

2.1. Basic Current Measurement

A simple structure capable of sub-threshold current measurement is a transistor in pass-gate configuration. In this setup, the source of the transistor is connected to a power rail (VDD or GND for PMOS and NMOS, respectively), drain connected to a common integrating dual-slope ADC [8] and gate controlled by an external bias, as shown in Figure 1. Integrating ADCs are highly accurate and conducive to current measurements by charging an integrating capacitor, although conversion times are slow. The resolution of the ADC in this design is configurable up to 13 bits, allowing a trade off between accuracy and measurement time.

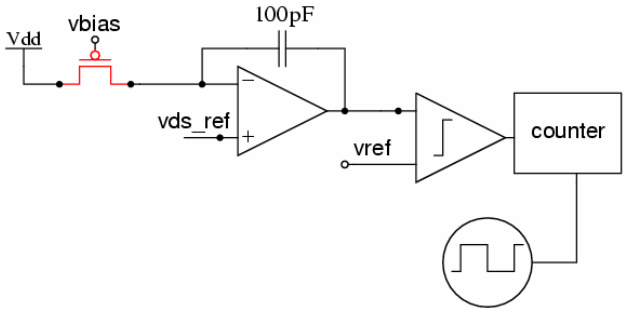

Figure 1: Simple I_D measurement scheme with integrating ADC

V_{GS} is controlled by an external V_{BIAS} input signal while V_{DS} is set by the operational amplifier through enforcement of a virtual ground between the positive and negative terminals of the amplifier. Using a dual-slope approach (extra control logic not shown) rejects amplifier gain errors as both the charging and discharging phases follow identical voltage paths.

2.2. DUT Array and Hierarchical Access Scheme

Requiring an ADC per DUT would consume valuable chip real-estate, as well as lead to concerns regarding variation in the measurement circuits affecting the accuracy of the computed deltas between each DUT. Instead, a hierarchical multiplexing scheme is used such that all DUTs share common measurement circuitry. The multiplexing scheme is organized such that there are 128 NMOS and 128 PMOS DUTs per "bank" (Figure 2) with all banks connecting to the *p_grid* and *n_grid* grids. These grids subsequently connect to a single integrating ADC. Sharing a single replicate of the measurement circuitry ensures that any variation in this set of circuits is common to all DUTs and does not affect computed variances.

Each DUT has separate row and column enable transistors, making access analogous to a memory organization. Each bank also contains a low-leakage, high-V_T pass-gate and controlling logic to effectively isolate the bank from the measurement circuitry when the bank is not enabled. 540 banks are organized into six sections, resulting in a dense array of 127 columns x 90 rows x 6 sections = 68,580 transistors of each type in 2mm x 2mm. This is one of the largest and most dense array of DUTs compared to those found in the reference work provided in Section 1. Row, column and section-enable serial-shift registers control access to the DUT array.

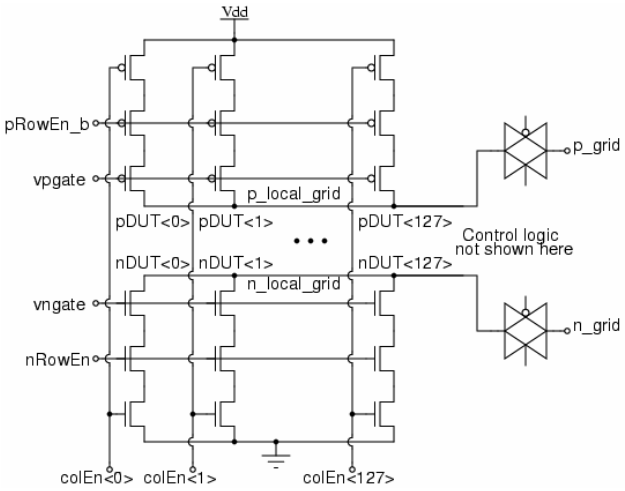

Figure 2: Simplified schematic of individual bank

While the row and column access transistors introduce resistance and variation, HSpice simulations have shown that a +/-10% variation in either L or V_T of the access transistors has <0.5% effect (Section 4.2) on I_{DS} of the DUT being accessed. Despite column access transistors being turned completely off for all other DUTs, a finite leakage current, I_{leak}, through the row and column access transistors and the "off" DUTs adds to the I_{DS} of the DUT being accessed. When I_{DS} of the accessed DUT is large, corresponding to a larger V_{GS}, I_{leak} is a negligible component and can safely be ignored. However, as will be described in Section 3, it is desirable to set V_{GS} as low as possible to benefit from the convergence of n at low V_{GS}. At gate biases below 0.25V, I_{DS} reduces to single nanoamps and lower, meaning that even small drain-source leakage currents and drain/source-body junction

currents accumulate over the "off" DUTs and their access transistors.

2.3. Mitigating "Off" Device Leakage and Junction Currents

The well-known stack effect present as a result of the row and column access transistors does reduce the drain-source leakage current [9]; however, it is not eliminated, nor does transistor stacking eliminate drain-substrate and source-substrate reverse-bias junction currents. While I_{leak} can be treated as an offset and subtracted from any current measurements made, offsets larger than the actual current being measured result in lowered measurement confidence. As a result, circuit or architecture-level techniques must be employed to mitigate this parasitic current contribution.

By splitting the array into hierarchically accessed banks as described in Section 2.2, I_{leak} is reduced as the number of transistors directly connected to measurement circuitry is reduced to the DUTs in a single bank (128) plus the pass-gates of each bank (540). The effect of extra resistance in the current measurement path as a result of the added pass-gate is small, as the measured currents are low (<1μA). Since the bank pass-gates are high-Vt, dual-gate-oxide, low-leakage devices, the majority of the parasitic current contribution is due to the 127 "off" DUTs in the bank being accessed. The choice of 128 DUTs, of each type, per bank is not arbitrary; rather, it is a compromise between having the fewest additional transistors per DUT and attaining low I_{leak}. With 128 DUTs per bank, I_{leak} is approximately on the order of single nanoamps.

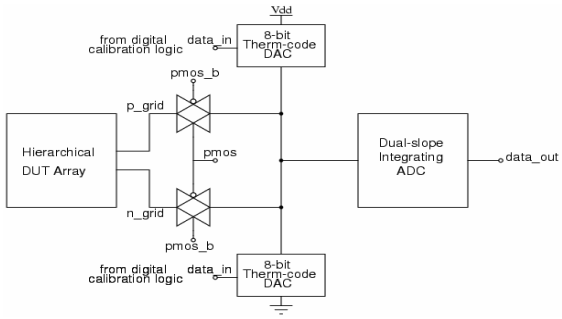

Figure 3: Simplified Top-Level Block Diagram

To further decrease I_{leak}, an active current subtraction scheme is employed. As shown in Figure 3, two (source and sink) 8-bit thermometer-code DACs [8] with digital control logic are added to actively add or subtract an amount of current equivalent to $-I_{leak}$. The digital logic implements a binary search algorithm that uses the output of the ADC to converge upon the correct DAC input value and acts as an auto-zeroing mechanism. For example, when trying to measure the first NMOS DUT in a bank, the auto-zeroing will first be run when all DUTs in the bank are off. If the ADC output is anything but 0 after the first auto-zeroing measurement, the digital logic will completely turn on one of the two DACs shown in Figure 3 in response to the direction of I_{leak}. If I_{leak} is being drawn from the measurement circuitry to ground, the algorithm will turn on the source DAC (top of Figure 3) in order to "source" I_{leak} and remove its effect from the measurement. Analogously, the algorithm will turn on the sink DAC (bottom of Figure 3) to "sink" an I_{leak} flowing from V_{DD} to the measurement circuitry. Subsequent measurements are used to refine the DAC control word in a logarithmic fashion. Auto-zeroing is performed once for each bank being tested at a specific gate bias. Due to the discrete nature of a DAC as well as limited resolution, the auto-zeroing will not be perfect and residual I_{leak} is treated as an offset and subtracted from DUT current measurements.

3. Isolation of ΔV_T

In order to isolate the measurement of ΔV_T, we take advantage of the sub-threshold regime of operation. In this regime, the drain current of a transistor can be expressed by Eq. 1.

$$I_D = I_o \cdot e^{\frac{V_{GS} - V_{To} - (\gamma' \cdot V_{SB}) + \eta \cdot V_{DS}}{n V_{th}}} \cdot \left(1 - e^{\frac{-V_{DS}}{V_{th}}}\right) \qquad \textbf{Eq. 1}$$

$$n = \frac{\Delta V_{GS}}{V_{th} \cdot \Delta \log(I_D) \cdot \ln(10)} \qquad \textbf{Eq. 2}$$

where I_o is the drain current at $V_{GS} = V_{To}$, γ' is the body-effect coefficient, η is the Drain-Induced Barrier-Lowering (DIBL) coefficient, n is the sub-threshold slope ideality parameter as defined by Eq. 2, and V_{th} is the thermal voltage.

We immediately note that the (*1-e*) multiplicative term only results in a significant multiplicand if V_{DS} < ~0.2V at room temperature, 25C. It will be shown later that to minimize the V_{DS} contribution in the main exponential term, V_{DS} is minimized but kept >0.2V. V_{SB} is set to 0V in order to eliminate any body-effect on V_T. Eq. 1 can now be simplified to:

$$I_D = I_o \cdot e^{\frac{V_{GS} - V_{To} + \eta \cdot V_{DS}}{n V_{th}}} \qquad \textbf{Eq. 3}$$

Assuming for two identical devices that nothing except V_T varies, the natural logarithm of the ratio of the device currents multiplied by a scale factor will result in ΔV_T as shown in the following equations:

$$\ln\left(\frac{I_{D2}}{I_{D1}}\right) = \ln\left(\frac{I_o \cdot e^{\frac{V_{GS} - V_{To2} + \eta \cdot V_{DS}}{n V_{th}}}}{I_o \cdot e^{\frac{V_{GS} - V_{To1} + \eta \cdot V_{DS}}{n V_{th}}}}\right) \qquad \textbf{Eq. 4}$$

$$\ln\left(\frac{I_{D2}}{I_{D1}}\right) = \ln\left(e^{\frac{V_{To1}-V_{To2}}{nV_{th}}}\right) \qquad \textbf{Eq. 5}$$

$$nV_{th} \cdot \ln\left(\frac{I_{D2}}{I_{D1}}\right) = \Delta V_{T1,2} \qquad \textbf{Eq. 6}$$

While it was shown in Section 2.1 that V_{DS} is controllable by circuit setting, variation will mean that n will likely vary for each device as well. Taking this into account, the above equations can be reworked and shown to provide the relationship in Eq. 7 between the drain currents, threshold voltages and n.

$$n_1 V_{th} \cdot \ln\left(\frac{I_{D1}}{I_{D2}}\right) - \frac{(n_2 - n_1)}{n_2} \cdot V_{GS} = \frac{n_1}{n_2} \cdot V_{T2} - V_{T1} \qquad \textbf{Eq. 7}$$

Since Eq. 7 provides no simple, analytic solution for ΔV_T, the values of of n_1, n_2, V_{GS} and at least one device's V_T must be known to compute the other's and thus a delta between the two. Eq. 2 shows that n can easily be computed using two measurements of I_D at differing values of V_{GS}. We also note that by using a small value for V_{GS} we can minimize the contribution of the second term in Eq. 7 in two ways: a smaller V_{GS} results in 1) a smaller multiplicand, and 2) as can be seen in Figure 4, the instantaneous value of n tends to converge at extremely low V_{GS} despite variation in V_T, allowing use of Eq. 6 rather than Eq. 7.

Ascertaining the value of one of the device's threshold voltage is more difficult, but possible by measuring the sub-threshold drain current at numerous values of V_{GS} and fitting the data using Eq. 1 to find a value of V_T. This known V_T can then be used to compute the V_T for every other DUT. By using Eq. 7, along with two additional measurements for each DUT (to compute n), a complete V_T map can be ascertained.

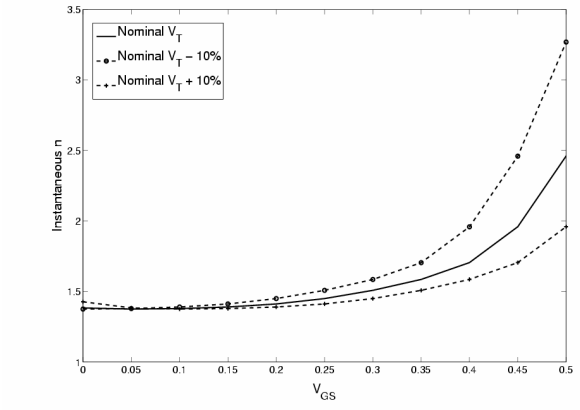

Figure 4: Convergence of n at low V_{GS} despite V_T variation

4. Simulation Results and Test Chip

All circuits described in the previous section were simulated in Hspice using model files from National Semiconductor's 0.18µm process. The following subsections detail results of these simulations. These results were used to implement a test-chip in the same process to test the aforementioned circuits.

4.1. V_T Isolation Results

To test the theory set forth in the Section 3, simulations were performed in which a single DUT, with row and column access transistors, charges an integrating capacitor. Since the operational amplifier forces a virtual ground at the inputs, V_{DS} remains constant as the capacitor is being charged and I_{DS} can then be solved for using Eq. 8. Simulations were performed with DUT V_T and channel length being varied by +/-10% and V_{DS} being varied across the allowable range, determined by the output stage of the operational amplifier. The range in this design is 0.3V – 1.5V.

$$I = \frac{C \cdot \Delta V}{\Delta t} \qquad \textbf{Eq. 8}$$

The plots in Figure 5 show the simulation results. The top plot in the figure varies DUT V_T and V_{DS}, while the bottom plot varies DUT channel length and V_{DS}. Both plots show the relative change in current from the nominal V_T or L at a given value of V_{DS}. The plots clearly show that the arrangement detailed in Section 2.1 is more sensitive to changes in I_{DS} as a result of V_T variation rather than L variation, especially at low V_{DS}. These results are consistent with the theory outlined in Section 3.

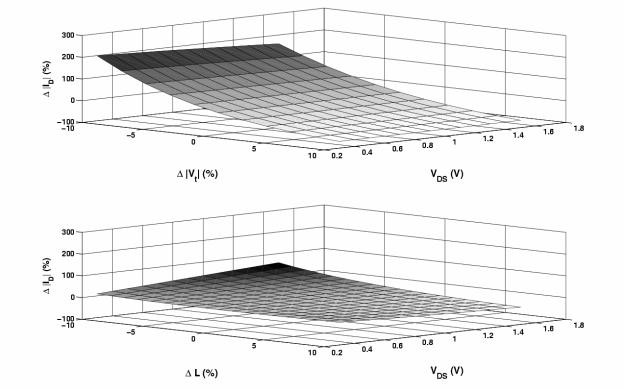

Figure 5: Percent change in current from nominal value at a given V_{DS} when varying V_T (top plot) and channel length, L (bottom plot)

To quantify these results further, the sensitivity of I_{DS} to either ΔV_T or ΔL can be computed by taking the derivative with respect to ΔV_T and ΔL, respectively.

Taking the ratio of these derivatives gives the sensitivity ratio of the circuit to ΔV_T and ΔL. Since it is clear from Figure 5 that the circuit is least sensitive to ΔL at low values of V_{DS}, these derivatives are only calculated for the lowest V_{DS} value allowable, 0.3V. Figure 6 plots the ratio of $(\partial I_D / \partial L)$ to $(\partial I_D / \partial V_T)$ for V_{GS} ranging from 0.35V to 0.5V. Lower values of V_{GS} were not plotted, as a trend in decreasing sensitivity to ΔL with larger values of V_{GS} is evident from the figure. However, Section 3 discussed employing lower values of V_{GS}, where the value of n converges despite variation. The results of sensitivity analysis imply that simply measuring the value of n, as discussed in Section 3, and using a V_{GS} near the nominal V_T for the process provide more benefit in extracting ΔV_T than attempting measurements at extremely low gate biases.

For all values of V_{GS} plotted, the sensitivity ratio through the majority of the variation range is below 0.1, meaning the circuit is at least 10X more sensitive to V_T variation than it is to L variation. Furthermore, sensitivity to V_T variation peaks in the vicinity of the process' nominal values of V_T and L. Since variation in these parameters are typically normally distributed about the nominal value, the majority of variation measured will be in the high-V_T-selectivity region of operation, giving high confidence that the measured variation is primarily a result of threshold voltage variation.

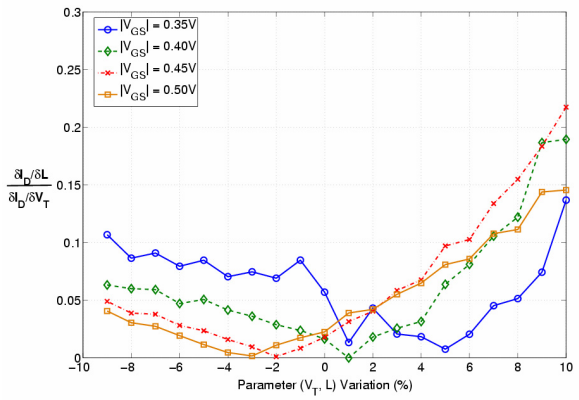

Figure 6: Circuit Sensitivity Ratio

Simulations were performed where a transistor was subjected to variation (V_T or L or both) and the ADC outputs used to determine the amount of variation with both Eq. 6 and Eq. 7. All simulations were done with $|V_{GS}|$ = 0.35Vand $|V_{DS}|$ = 0.3V and the ADC resolution set to 10 bits. Simulations were also done at $|V_{GS}|$ = 0.345V, the results of which used in conjunction with Eq. 2 to calculate n. Table 1 contains the results of these simulations. It should be noted that the simulations where only V_T is varied result in approximately a 10% error in the extracted deltas, which is primarily a result of inaccuracy in extracting n. This alone would indicate that this test-structure can resolve deltas of approximately 1% of the nominal V_T. However, since the sensitivity of the circuit to V_T variation is not infinite, resolution is reduced to approximately 2% of nominal V_T. Resolution can be enhanced by increasing ADC resolution and more measurements at differing values of V_{GS} to extract n.

Variation Type	Extracted ΔV_T
+10% V_T	+10.9%
-10% V_T	-11.0%
+10% V_T, +10% L	+11.9%
+10% V_T, -10% L	+9.3%
-10% V_T, -10% L	-12.0%
-10% V_T, +10% L	-11.3%
+3% V_T, +3% L	+3.6%
-3% V_T, -3% L	-3.1%

Table 1: Extracted V_T variation vs. subjected variation

4.2. Access Transistor and Resistance Effects

In order to hierarchically access a large number of DUTs within an array, each DUT requires row and column access transistors, and each bank of DUTs requires a pass-gate. These devices introduce additional resistance, potentially lowering the V_{DS} and corresponding I_D of the DUT due to the finite R_{out} of the devices. However, operation in the sub-threshold regime produces small currents which are not perturbed much by even fairly large resistances. Simulations were carried out to evaluate this impact. The test circuit from Section 2 was used in these simulations with and without row, column and bank access transistors. Table 2 shows that the impact of these transistors and variation within them is less than 0.5% of the simulated current without any access transistors.

Test Scheme	Relative Difference
DUT w/o access transistors	0.00%
DUT w/access transistors	-0.43%
DUT with -10% ΔL in access transistors	0.26%
DUT with -10% ΔV_T in access transistors	0.23%

Table 2: Simulated current differences due to inclusion and variation within access transistors ($|V_{GS}|$ = 0.35V and $|V_{DS}|$ = 0.3V)

Another possible source of inaccuracy in implementing a large array is variation in the distance current must travel to the measurement circuitry, resulting in different resistances seen by each DUT to the ADC. However, simulations show that even with a 1kΩ

difference in resistance, the relative current difference is only 0.1%. Furthermore, the path from each DUT bank to the ADC is implemented as a dense metal grid spanning multiple metal layers to provide the lowest possible resistance. Process data and simulations indicate that a minimum width wire spanning 2mm has a resistance of approximately 500Ω. However, the grid is implemented with many 3X minimum-width wires spanning four metal layers, decreasing the overall resistance substantially. Additionally, the resistance difference between any two DUTs cannot be more than the resistance of a single minimum-width wire spanning the entire array, so we conclude that resistance variations in the grid will have negligible effect on measured currents.

4.3. Test Chip

A test chip has been implemented and submitted for fabrication on National Semiconductor's 0.18µm process. The test chip occupies 3.1mm x 2.6mm of area, with 2mm x 2mm dedicated to the DUT array (bright white square in Figure 7). The circuitry on the middle right of the figure comprises the integrating capacitor, ADC, DACs and digital control logic. Blank areas are filled with dummy fill (not shown) for improved CMP planarity. Fabricated chips are expected in January 2007.

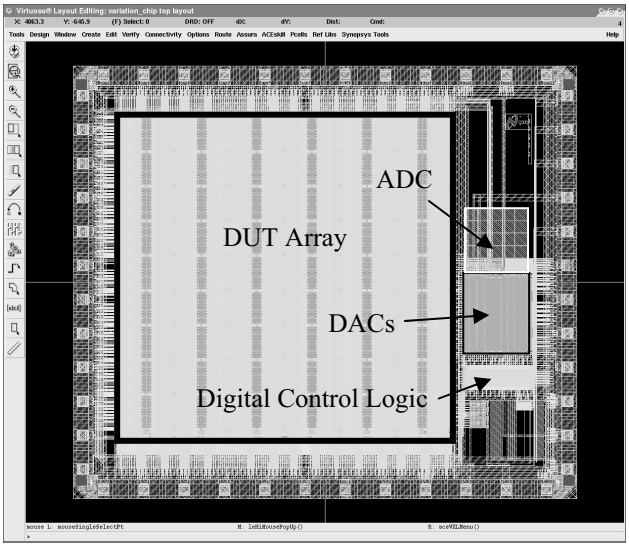

Figure 7: Chip Layout

5. Conclusions and Future Work

We have developed a method to isolate and extract threshold voltage variation efficiently in large transistor arrays. By measuring device currents when biased in the sub-threshold regime, where current is exponentially dependent on $V_{GS} - V_T$, simulations have shown that the presented test structure is at least 10X more sensitive to variation in threshold voltage than it is to variation in channel length. Hierarchical access and auto-zeroing circuitry allow for a large number of devices (>100K) to share the same measurement circuitry with little reduction in measurement accuracy.

Upon successful testing of the test structure on the submitted 0.18µm test chip, we plan to scale the chip to advanced technologies such as 45nm bulk CMOS or possibly FinFET processes to evaluate threshold voltage variation on undoped, thin-body processes.

Acknowledgments

This work has been supported in part by the MARCO Center Circuits and Systems Solutions (C2S2). We thank National Semiconductor for access to fabrication and technology design data.

References

[1] S. R. Nassif, "Design for Varaibility in DSM Technologies," *Proc. ISQED*, pp. 451-454, 2000.

[2] X-W. Lin, "Design and Process Variability – the Sources and Mechanisms," *DAC Tutorial – Practical Aspects of Coping with Variability: An Electrical View*, 2006.

[3] X. Xie, D. Boning, F. Meyer, and R. Rzehak, "Analysis of Nanotopography and Layout Variations in Patterned STI CMP," *Int. Conf. on Planarization/CMP Technology*, 2006.

[4] J. S. Panganiban, *A Ring Oscillator Based Variation Test Chip*, Masters of Engineering Thesis, MIT EECS, 2002.

[5] L.-T. Pang and B. Nikolic, "Impact of Layout on 90nm CMOS Process Parameter Fluctuations," *IEEE Symposium on VLSI Circuits Technical Digest,* 2006.

[6] M. Bhushan, M.B. Ketchen, S. Polonksy and A. Gattiker, "Ring Oscillator Based Technique for Measuring Variability Statistics," *Proc. Of the Int. Conf. on Microelectronic Test Structures*, pp. 87-92, 2006.

[7] V. Wang and K. L. Shepard, "On-Chip Characterization Macro for Variability Analysis," Submitted to *ISQED*, 2007.

[8] D. A. Johns and K. Martin, *Analog Integrated Circuit Design*, John Wiley & Sons, Inc., New York, 1997.

[9] Z. Chen, M. Johnson, L. Wei, and K. Roy, "Estimation of Standby Leakage Power in CMOS Circuits Considering Accurate Modeling of Transistor Stacks," in *Proc. Int. Symp. Low Power Electronics and Design*, pp. 239-244, Aug. 1998.

[10] K. Terada and M. Eimitsu, "A Test Circuit for Measuring MOSFET Threshold Voltage Mismatch," *Int. Conf. on Microelectronic Test Structures*, 2003.

[11] K. Agarwal *et al.*, "A Test Structure for Characterizing Local Device Mismatches, *Symp. On VLSI Circuits*, 2006.