The Report Committee for Mayank Gupta
Certifies that this is the approved version of the following report:

Design and Implementation of a Scribe Line Measurement Transistor Test Array Structure in 14nm FinFET CMOS Technology

APPROVED BY
SUPERVISING COMMITTEE:

Supervisor:

Jacob A. Abraham

Hoang Dao

Design and Implementation of a Scribe Line Measurement Transistor Test Array Structure in 14nm FinFET CMOS Technology

by

Mayank Gupta, B.S.E.E

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Engineering

The University of Texas at Austin

May 2015

Dedication

This work is dedicated to my family, Kate, Priya, Ravi, and Nico.

Acknowledgements

The author would like to acknowledge the following people for their contribution, mentorship, advice, and encouragement:

- My mentor, Hoang Dao, at AMD for his coaching and giving me the opportunity to work on this project.
- My supervisor, Dr. Jacob A. Abraham, at the University of Texas at Austin for his guidance throughout the class.
- My manager, David King, and my co-workers in the silicon enablement team at AMD.
- The layout team under Roman Guzman, including Karim Fazlul, Ellen Wilkosz, Shajib Kamruzzaman, and Syeda Afrin at AMD.
- The technology team including Ioan Cordos, Kelvin Holub, and Da Zhang at AMD.
- The legal and public relations teams including Karen Prairie, Kim Vo, Sarah Blankenship, and Cheryl Heinsohn at AMD.

Abstract

Design and Implementation of a Scribe Line Measurement Transistor Test Array Structure in 14nm FinFET CMOS Technology

Mayank Gupta, M.S.E.

The University of Texas at Austin, 2015

Supervisor: Jacob A. Abraham

Submicron fin-shaped field effect transistor (FinFET) process technologies pose a variety of challenges for foundries ramping designs into production due to parameter variation. Accurate and relevant electrical transistor parametric data are required to correlate product yield and performance to the process technology throughout the manufacturing cycle. Typically, transistor data are obtained from foundry-designed and isolated transistors test structures which do not resemble the layout, physical attributes, or design style of the transistors in the standard cells of the actual design. As a result, the topographical, density, and stress effects on the transistors of a custom or synthesized design layout are not reflected in the test measurements. To address this issue, a scribe line macro (SLM) test structure with modular transistor arrays utilizing standard cells from the design library was designed for a 14nm FinFET complementary metal oxide semiconductor (CMOS) technology process. The SLM test structure fits in a 52μm by 2000μm footprint and can be printed at multiple locations in the scribe lines of a 300mm wafer. The SLM test structure comprises of four n-type metal oxide semiconductor (NMOS) and four p-type

metal oxide semiconductor (PMOS) transistor arrays containing 240 FinFET transistors of 33 types that can be tested rapidly and in parallel. The SLM test structure is designed to generate silicon data for parametric transistor measurements such as I_{DSAT}, I_{DLIN}, I_{DOFF}, V_{TSAT}, and V_{TLIN} using a parametric inline tester in the manufacturing line of a foundry. These silicon measurements will be correlated to simulation data extracted from the netlist of the SLM test structure to eventually improve the 14nm FinFET technology libraries and silicon models. The SLM test structure design will be continuously improved for future technology processes with a goal of enabling accurate and design-relevant transistor parametric measurements to help drive cycle time, cost, and yield improvements.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction and Background

Submicron CMOS technology scaling poses a variety of challenges for the semiconductor industry. For planar CMOS technology, these challenges include manufacturing and lithography precision issues, process variability, and the performance degradation due to the physical and electrical limitations of the transistor in smaller geometries. Along with yield and cost, the industry has strived to maintain or improve CMOS transistor device performance metrics with each process node. The key transistor metrics being higher drive current (high I_{DSAT}), lower off-current and leakage (low I_{OFF}), lower delay (low capacitance and series resistance), and better control (precise targeting and low variation) of the transistor. Until the 20nm planar CMOS technology node, these improvements were attainable via transistor scaling and various manufacturing and process advancements, such as strained channel for increased mobility, better dopants, high K metal gates (HKMG), use of dielectrics, ultra-violet lithography techniques, etc. [1]. Beyond the 20nm planar CMOS technology node newer challenges arose which required new approaches, architectures, and techniques. Some of these newer challenges include reduced mobility and velocity saturation of the transistors, higher leakage due to relatively higher voltage supply and shorter gates, increased capacitance and higher resistances in the channels, increased process variability due to random dopant fluctuations, line edge roughness and spatial effects, feature size lithography limitations, and increased transistor reliability issues [2]. Along with advances in lithography and design for manufacturing (DFM) techniques, FinFET transistors emerged as a solution to some of the issues with sub-20nm planar CMOS transistor scaling.

FinFET transistors provide a new type of self-aligned multi-gate device in a conventional CMOS process with a quantized channel width [3]. This allows for several

improvements over planar CMOS devices, such as improved gate control due to steep sub-threshold region slope and reduced short channel effects, lower leakage and power due to a thinner body region and lower V_T, improved mobility, lower parasitic capacitance, improved area efficiency, and improved transistor V_T matching [4]. These improvements will allow CMOS technology scaling to extend for several more generations beyond the 20nm node. Despite these advantages FinFET devices pose additional variability and process challenges due to higher parasitic capacitance between gates and the top/bottom of the source/drain, high access resistance due to a thinner body, performance sensitivity to fin width, and V_T adjustment difficulties [2]. Due to its 3D structure, FinFET device manufacturing also requires unique lithography techniques for gate patterning and tighter process control to print features down to 10nm and smaller in length [3].

These issues of process and parameter variability essentially translate to higher design, manufacturing, and test cost. Increased parameter variation makes the design process a more probabilistic endeavor requiring sophisticated models, more margin and redundancy, and costly area and power trade-offs. Process variation across devices and spatially across a wafer can cause significant variation in overall chip performance and yield [5]. This makes it crucial for foundries to monitor AC and DC performance of transistors in the manufacturing line to optimize the process to meet yield and performance targets as well as to generate data to improve device modeling [6]. This monitoring or inline testing is done by probing scribe line macro (SLM) device test structures on wafers using parametric testers at multiple steps in the manufacturing line. These structures are named as such because they are placed in the scribe lines of a wafer (i.e. between functioning die).

This paper presents the design and implementation of a FinFET SLM transistor test structure that meets the primary need for accurate and continuous inline monitoring of FinFET parameter variation in a manufacturing line. The rest of this chapter provides an

overview of inline testing, a background of SLM test structure design and their shortcomings, and a discussion of an SLM FET array design which forms the basis of the FinFET SLM test structure presented in the paper. A high-level overview of the FinFET SLM test structure and its improvements over a typical SLM design are also presented at the end of this chapter.

1.1 INLINE TESTING OVERVIEW

Inline testing is the measurement of various parameters of test devices, such as CMOS transistors, FinFET transistors, interconnects, resistors, ring oscillators, latches, etc., in an SLM test structure placed in multiple locations between the functioning die on a wafer. This testing is accomplished using standard parametric testers in a semiconductor manufacturing line. SLM structures are designed to be long, skinny, compact, product representative, and easily placeable in the scribe lines of the wafer with no impact to die area [6]. Typical SLM structures consist of isolated devices or test structures with inputs and outputs connected to contact pads. Certain SLM structures may utilize an addressable array of test structures designed to extract DC measurements. Such SLM array structures, usually for transistor measurements, are self-timed, self-calibrating, and function with DC inputs from parametric testers to enable rapid testing with minimal post-processing of data [6]. Certain SLM structures utilize scanable latches or registers and have circuitry to support AC or high-speed bench testing, although this type of inline testing is slower and done infrequently [6]. SLM structures are designed to share a limited number of pads and utilize Kelvin connected test structures to maintain measurement accuracy [7]. The devices under test (DUTs) in the SLM structure typically utilize no more than four metal layers so they can be tested early in the manufacturing line to provide feedback. To accurately

characterize parameters across all corners of a technology process and design, the SLM test structure may contain multiple identical devices over a range of device sizes, widths, V_T types, features, number of fingers, diffusion profiles, and numerous other physical or electrical attributes [7]. Because the DUTs must be immune to statistical fluctuations of single devices, measurements are usually averaged over a large number of identical DUTs across the SLM and across the wafer [6]. The measurements may also require data calibration and error correction.

Inline testing using SLM test structures provides feedback of device parameters for design modeling, yield and performance prediction, and process tuning to ultimately reduce die cost associated with variation. This learning is now more important than ever as we scale into submicron FinFET CMOS technology nodes, where the growing impact of transistor parameter variation (i.e., whether deterministic, systematic, or random) necessitates a design-technology co-optimization strategy [8]. For example, due to the constraints of deep sub-wavelength lithography, foundries must start early and continuously evaluate FinFET transistor parameters to optimize the gate pitch of devices to positively affect the design choices for subsequent revisions of a product [8]. Another example is the need to correctly model and control FinFET fin width variation and discrete tuning which affects circuit performance. Yet another example is the targeted evaluation of patterning in FinFET standard cells, made possible by submicron lithography, using inline testing data [8]. Such design-technology co-optimization approaches demonstrate the need for inline monitoring of devices that are as close to the standard cell design implementation as possible for optimal design modeling.

Typically, SLM test structures for transistor measurement are designed by foundries and use isolated transistors or passive devices which may not resemble a true design layout. This will become an issue as we move into submicron FinFET technology

nodes where inline testing data from foundry-designed SLM structures will not represent actual on-die device parameters due to the increased effects of process variability. SLM DUTs will need to be design-like standard cell implementations in order to do the following: to provide accurate parametric transistor data for variation design modeling, to accurately predict on-chip performance and yield metrics, to help evaluate lithography and DFM schemes utilized on chip, to mimic stress from surrounding cells as they would exist in a standard cell layout, and to mimic the density and topological effects due to surrounding cells. SLM DUTs must utilize as many multiple devices flavors as are used in the design to give the best device matching data as possible.

The FinFET SLM transistor test structure design presented in this paper meets this need by using a product-like design flow with minimally modified standard cells and filler cells from the design standard cell library. The FinFET SLM test structure features a true design-like implementation of DUTs in an easily testable, addressable array configuration.

1.2 SLM FET Array Design

The FinFET SLM test structure is based on an existing SLM FET array design that describes an addressable array structure to rapidly collect MOSFET parameters in a compact configuration suitable for early process learning. The design uses a modular approach for test structure integration that supports parallel testing [9]. The modular arrangement favors a design of experiment (DOE) approach which allows for a large number of different MOSFET transistors or DUTs to be placed easily and tested rapidly to extract a full range of I-V characteristics including leakage currents in the sub-threshold region [9]. The design of the SLM FET array consists of 8 array units: four NMOS arrays and four PMOS arrays. Each array unit consists of 30 DUTs, circuitry for supplying

currents and voltages to the terminals of the DUTs, and a 5-bit decoder to select one, all, or none of the 30 DUTs in an array as represented in Figure 1.1 [9]. Each array unit is placed between two contact pads which provides a shared connection for all 30 DUTs of an array unit. The source and drain of all DUTs are connected to the contact pads while the gate connections are activated by the decoder circuit.

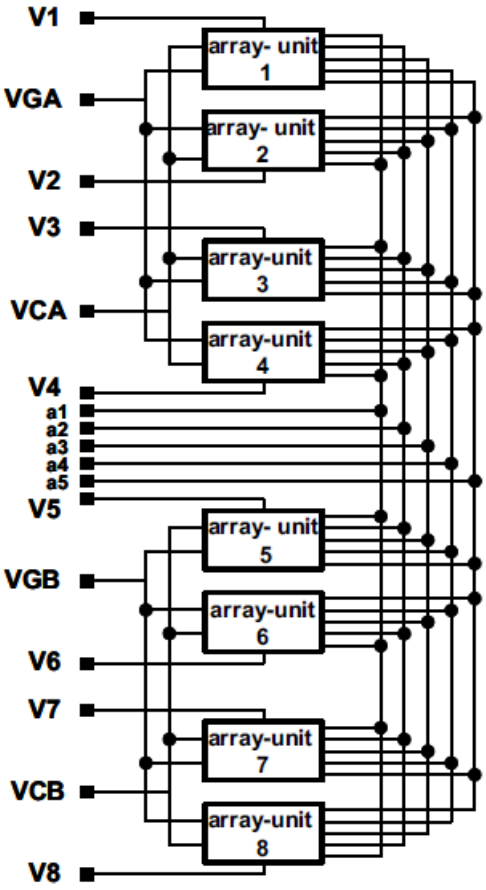

Figure 1.1: Top level schematic of SLM array [9].

In Figure 1.1, the V1 through V8 signals are the drain connections for each array unit and can be used to set an appropriate drain-to-source voltage for the selected DUT of that array

unit. Supplies named VCA, VCB, VGA, and VGB used to set the gate voltage of the selected DUT of an array unit. VGA and VCA are dual supplies to the NMOS arrays where VGA is used to turn on or select a DUT, while VCA is used to turn off or unselect a DUT. The VCA and VCB supplies serve the same function for the PMOS arrays. In Figure 1.1, the a[5:1] inputs are the decoder bits that are shared among all array units and can be encoded with b'00000 to select no DUTs, b'11111 to select all DUTs, or an appropriate 5-bit binary encoding to select a single DUT from 1 through 30 for a given array unit. An array unit is activated by one of the V1 through V8 drain signals while one, none, or all DUTs of the activated array unit is selected by the a[5:1] signals which steers the VGA and VGB supplies to the gates of the selected DUTs and VCA and VCB supplies to the gates of the unselected DUTs. The multi-port design of the SLM FET array allows parallel testing of multiple array units for rapid data collection. The modular array design supports the implementation of multiple types of DUTs across device type, dimensions, and layout styles [9]. It also supports the placement of identical types of DUTs across array units to model random statistical variation across wafer to help identify systematic process variations [9].

A 3x10 matrix DUT design of an NMOS array unit is depicted in Figure 1.2. Each DUT in the 3x10 NMOS array unit in Figure 1.2 is connected to a shared source and drain contact pad on either side to minimize parasitic voltage drop across the DUT [9]. Vertical metal wires run along the length of the SLM and tap into each array unit. These metal wires are the VDD and GND supplies for the decoder and array unit, decoder inputs, and low current gate voltage inputs (VGA, VCA, VGB, and VCB) to the array unit via the decoder. The decoder inputs from the pads are designed to source very low current and not impact DUT drain current measurements once the signals settle after a selection [9].

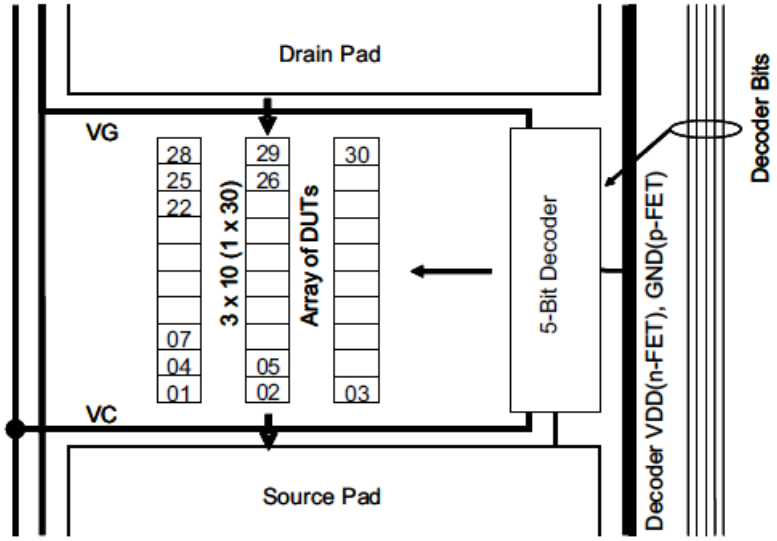

Figure 1.2: NMOS array unit physical representation [9].

The decoder circuit steers the dual gate voltage signals via low leakage bias circuits to select and unselect appropriate DUTs in an array unit. The authors suggest the use of low leakage pass transistors in series to minimize drain-to-source leakage currents in the bias circuits to keep the voltage drop on the gate voltage buses (VGA, VCA, VGB, and VCB) as low as possible [9].

Careful design of the decoder bias circuits allows the measurement of drain-to-source current I_{DS} for a selected DUT with minimal error as the remaining DUTs in the array remain unselected and are turned off. However, to measure the I_{DS} of the selected DUT in the sub-threshold region a correction must be applied for which additional measurements are needed [9]. Under the conditions that all DUTs of the array unit have similar I-V curves, as in the case where all DUTs in the array are identical, the I_{DS} of a selected DUT k over the entire sub-threshold region can be given by the following equation [9]:

$$I_{k\_VG} = \frac{\left(I_{k\_VG\_VC} - I_{all\_VC}\right)}{(1 - \eta_{all})}$$

In this equation, $I_{k\_VG\_VC}$ is the measured I_{DS} of a DUT k turned on with the remaining 29 DUTs of the array turned off, I_{all_VC} is the measured I_{DS} of all 30 DUTs turned off, and $\eta_{all} = I_{all\_VC}/I_{all\_VG}$ where I_{all_VG} is the measured I_{DS} of all 30 DUTs turned on [9]. The threshold voltage V_T of a DUT can also be determined by recording the gate voltage value applied to achieve a fixed drain current equal to I_{SV} * W/L, where W and L are MOSFET widths and channel length respectively, and I_{SV} is a value already defined for a given technology node and is usually between 20nA to 300nA for submicron CMOS technology nodes [9]. The SLM FET array is thus an efficient and adaptable inline testing solution for MOSFET transistors.

1.3 FinFET SLM Test Structure

The FinFET SLM test structure presented in this paper borrows the architecture and functionality of the SLM FET array design described above but ported entirely to a 14nm FinFET CMOS process and designed to fit within a 52µm by 2000µm footprint for a 1x25 pad SLM design. While the SLM FET array design described above extends to only one metal layer, the FinFET SLM test structure design in the paper goes up to the fourth metal layer (M4). Special design consideration was taken to minimize effective resistance due to additional metal layers. The FinFET SLM test structure will be placed in the scribe lines of 300mm production wafers and will be continuously monitored in the manufacturing line using Agilent 4082 parametric testers at M4 and final wafer electrical test (WET) insertions. The Agilent 4082 testers will provide power and ground supplies and test signals via special probe cards that are already designed to support the 1x25 pad SLM design. Test

algorithms to support parallel testing of the FinFET SLM test structure across multiple parameters are also defined.

The FinFET SLM test structure presented in this paper is an improvement over typical foundry-designed SLM structures, as it fulfills two key needs of FinFET CMOS semiconductor manufacturing. The first is the need for rapid, accurate, and continuous inline monitoring of FinFET parameter variation in a foundry manufacturing line. The second is the need for SLM DUTs to be true design-like standard cell implementations to more accurately measure transistor parameters to predict chip yield and performance and to build better design models.

The rest of this paper starts with Chapter 2, which details the design and implementation of the FinFET SLM test structure along with a discussion of the floor-planning, circuit, and layout specifications, as well as the DUT DOE. Chapter 3 goes into HSPICE® simulation results of the design along with a full description of the test algorithms and test plan for inline testing. Chapter 4 concludes with a discussion of future work and future improvement ideas for the FinFET SLM test structure.

Chapter 2: FinFET SLM Test Structure Design and Implementation

This chapter details the architecture of the FinFET SLM test structure (henceforth referred to as the SLM test structure) followed by a discussion of the design and implementation approach with an emphasis on the layout of the DUT array units. The SLM test structure presented here is based on an existing design which was ported to a 14nm FinFET CMOS process with a completely new implementation of the array units. The main design goal was to create a modular and easily testable SLM structure with an extensive transistor DOE that is suitable for inline testing after the M4 layer of the wafer is fabricated in a manufacturing line of a foundry.

2.1 ARCHITECTURE

The SLM test structure supports the inline testing of 240 FinFET DUTs in a multi-port multi-array configuration. The function of the SLM test structure is to enable rapid measurement of a full range of DC I-V test parameters of all FinFET DUT transistors. The SLM test structure was designed to be compact enough to fit in the scribe lines of a 300mm wafer so as to not affect actual die area. The SLM test structure fits a 1x25 pad SLM design already supported by the inline testers of the target foundry and is 52µm x 2000µm in size. It consists of four NMOS array units and four PMOS array units each with 30 DUTs that share a common drain and source pad. Each array unit can be measured individually or in parallel. Each DUT is implemented as a FinFET transistor with shared drain and source connections for the array unit. In each array unit, a DUT can be selected individually or in parallel with the 29 other DUTs via a dedicated 5-bit decoder which turns on the gates of either a single selected DUT or all 30 DUTs, and turns off the gates of either the 29 remaining unselected DUTs or no DUTs. The 5-bit decoder input is shared between all 8

array units. Each array unit also has its own power, source, and ground supplies which are shared or separated depending on the array unit. To enable rapid and parallel testing the power supplies to all four of the NMOS array units are sourced by a single supply, while PMOS array units 1 and 2 and PMOS array units 3 and 4 each have their own shared supplies to provide an extra degree of freedom for measurements. Each array unit has dedicated drain supplies which are used to set the V_{DS} to initiate current flow through the drain-source terminals of either the selected DUT or all 30 DUTs of an array unit in parallel depending on the decoder input. The entire design is combinational with no sequential latch, state-machine, or flop elements. Full details of the design hierarchy, architecture, and DUT DOE are presented below. Details on the floorplan, circuit design, layout, and key design attributes of the SLM test structure are addressed in Section 2.2 of this chapter.

2.1.1 Module Hierarchy and Components

The SLM test structure followed a modular and hierarchical design flow to speed up and simplify verification, circuit design, and layout. The inputs, outputs, supplies, components, and functions of the top level module, as well as all sub-modules of the design, are described in the following sections.

AMDFETH1A

AMDETH1A is the top level module of the design. A block diagram of this module is depicted in Figure 2.1. As seen in Figure 2.1, the pin list for this module is as follows:

- A 5-bit input: A[5:1].
- Seven (7) power supplies: VDD_N1TO4, VSS_1, VSS_2, VSS_3, VNWELL, VDD_P34, and VDD_P12.

- Four (4) gate control supplies or inputs: VGN, VCN, VGP, and VCP.
- Eight (8) drain control supplies or outputs: DR_N1, DR_N2, DR_N3, DR_N4, DR_P1, DR_P2, DR_P3, and DR_P4.

Also, seen in Figure 2.1, are the sub-modules of AMDFETH1A that are as follows:

- An ESP1N_M1ARY_A block: M1_A.
- Eight (8) arrays: NMOS_Array_1, NMOS_Array_2, NMOS_Array_3, NMOS_Array_4, PMOS_Array_1, PMOS_Array_2, PMOS_Array_3, and PMOS_Array_4 marked as N1, N2, N3, N4, P1, P2, P3, and P4.

The function of the AMDFETH1A module is as described in Section 2.1. The power supply for the four NMOS array units is the shared VDD_N1TO4 supply. The power supplies for the PMOS array units 1 and 2 are VDD_P12 and VDD_P34 for PMOS array units 3 and 4. The remaining power supplies VSS_1, VSS_2, VSS_3 and VNWELL are used as ground and body supplies for all the gates and DUTs throughout the design. The inputs to AMDFETH1A are the two dual sets of gate control supplies, VGN and VCN for the NMOS array units and VGP and VCP for the PMOS array units, and the 5-bit decoder input A[5:1]. Typically, VGN and VGP gate supplies will be set to a voltage to turn on the DUT, while VCN and VCP gate supplies will be set to a voltage to turn off the DUT. The outputs of the module are the 8 drain control supplies DR_Nx and DR_Px (where x = 1, 2, 3, 4). During inline testing, once a voltage is applied to the drain control supply of an array unit the current sourced from these pins can be measured to reflect the I_{DS} measurement of the selected DUT. A key sub-module of the AMDFETH1A is the decoder ESP1N_M1ARY_A whose components and sub-modules are described in the next four sections. The array units and their components are described at the end of Section 2.1.1.

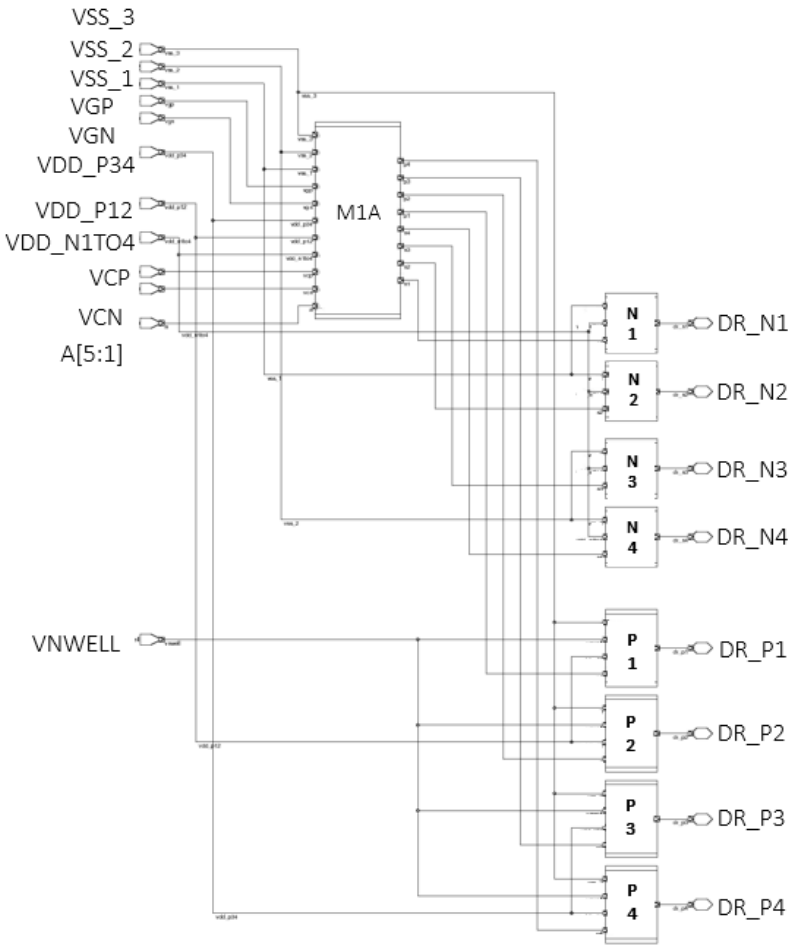

Figure 2.1: AMDFETH1A block diagram.

ESP1N_M1ARY_A

The ESP1N_M1ARY_A decoder sub-module is depicted in Figure 2.2. As seen in Figure 2.2, the pin list for this sub-module is as follows:

- A 5-bit input: A[5:1].
- Six (6) power supplies: VDD_N, VSS_N12, VSS_N34, VSS_P, VDD_P34, and VDD_P12.

- Four (4) gate control supplies: VGN, VCN, VGP, and VCP.
- Eight (8) 30-bit outputs: N1_sel[30:1], N2_sel[30:1], N3_sel[30:1], N4_sel[30:1], P1_sel[30:1], P2_sel[30:1], P3_sel[30:1], and P4_sel[30:1].

As seen in Figure 2.2, the sub-modules of ESP1N_M1ARY_A are as follows:

- Four (4) ESP1N_M1ARY_DECODE_P sub-modules: N1, N2, N3, and N4.
- Four (4) ESP1N_M1ARY_DECODE_N sub-modules: P1, P2, P3, and P4.

The function of the ESP1N_M1ARY_A sub-module is to generate the individual gate control signals for all 30 DUTs of all 8 array units based on the decoder input A[5:1] in parallel by steering the appropriate gate control voltages VGN, VCN, VGP, and VCP to the gates of the DUTs in all 8 array units. The gate control signals for all array units are the same since the decoder is shared across all array units. The decoding scheme for the outputs of this sub-module to an NMOS array unit is as follows:

- A[5:1] = b'00000 selects no DUTs and sets all Nx_sel[30:1] (where x = 1, 2, 3, 4) to VCN.
- A[5:1] = b'11111 selects all 30 DUTs and sets all Nx_sel[30:1] (where x = 1, 2, 3, 4]) to VGN.
- A[5:1] binary encoding selects a single DUT from 1 to 30 and sets the appropriate Nx_sel[i] (where x = 1, 2, 3, 4 and i = 1 to 30) output to VGN and the remaining 29 outputs to VCN.

The decoding scheme for the outputs of ESP1N_M1ARY_A to the PMOS array units is the same as the NMOS array units, except it steers the dual gate control supplies VGP and VCP instead. The sub-modules of the ESP1N_M1ARY_A and their components are described in the next three sections.

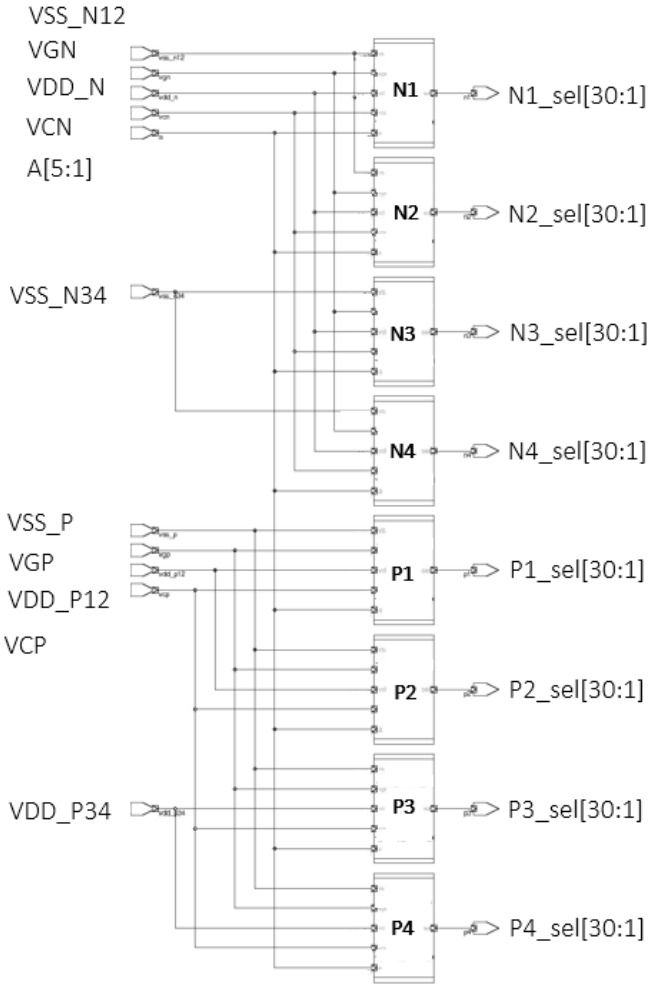

Figure 2.2: ESP1N_M1ARY_A block diagram.

ESP1N_M1ARY_DECODE_X

The ESP1N_M1ARY_DECODE_N and ESP1N_M1ARY_DECODE_P sub-modules are exactly the same and is denoted as ESP1N_M1ARY_DECODE_X. Figure 2.3 shows the pins and a subset of the sub-modules of ESP1N_M1ARY_DECODE_X. The pin list for this sub-module is as follows:

- A 5-bit input: A[5:1].

- Two (2) power supplies: VD and VS.
- Two (2) gate control supplies: VGN and VCN for the ESP1N_M1ARY_DECODE_N and VGP and VCP for the ESP1N_M1ARY_DECODE_P sub-module.
- A 30-bit output: sel[30:1].

As seen in Figure 2.3, the sub-modules of ESP1N_M1ARY_DECODE_X are as follows:

- An ESP1N_M1ARY_DECODE block: DEC
- Thirty (30) ESP1N_M1ARY_PASSGATE_X (where X = N or P depending on the type of the ESP1N_M1ARY_DECODE_X sub-module): X1, X2, X3, X4 etc.

Four of the 30 ESP1N_M1ARY_PASSGATE_X sub-modules (X1, X1, X3, and X4) are depicted in Figure 2.3. The remaining 26 sub-modules follow the same connection scheme and layout as shown for the four sub-modules in Figure 2.3, but are not shown in the interest of clarity. The function and decoding scheme of this sub-module is as described in the previous section, but only involves a single 30-bit output sel[30:1].

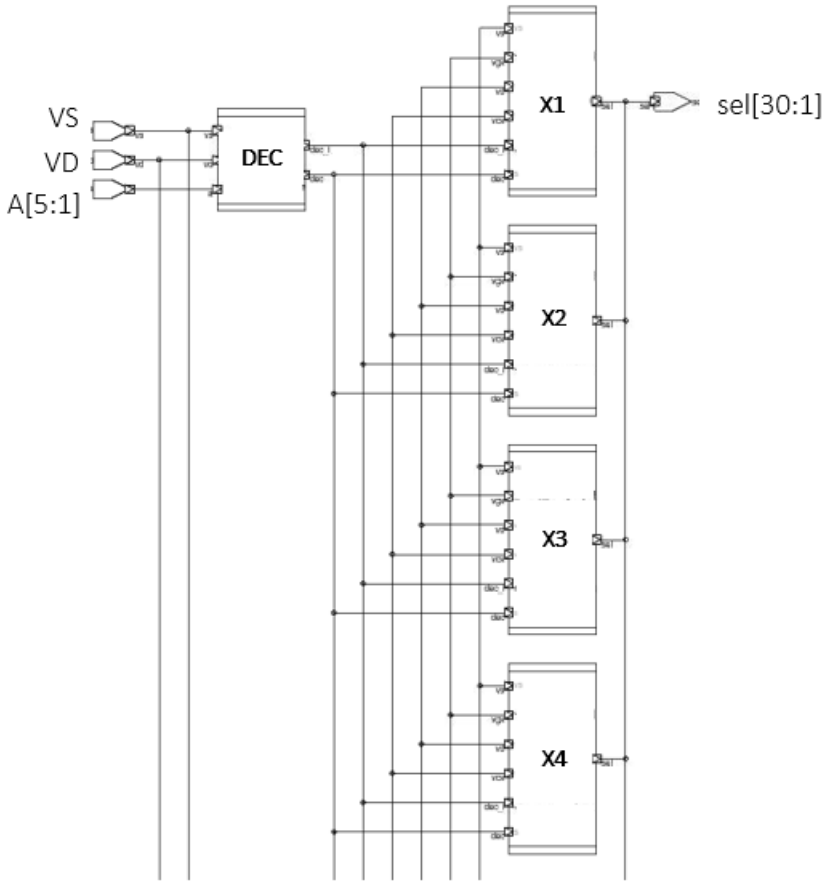

Figure 2.3: Partial ESP1N_M1ARY_DECODE_X block diagram.

ESP1N_M1ARY_DECODE

The ESP1N_M1ARY_DECODE sub-module is depicted in Figure 2.4. As seen in Figure 2.4, the pin list for this sub-module is as follows:

- A 5-bit inputs: A[5:1].
- Two (2) power supplies: VD and VS.
- Two (2) 30-bit outputs: dec_i[30:1] and dec[30:1].

The function of this sub-module is to perform the 5-bit to 30-bit 1-HOT decoding scheme described earlier for the ESP1N_M1ARY_A sub-module and is implemented using NAND gates and inverters. The power supplies VD and VS powers the source and body terminals of the transistors of the NAND and inverter gates. The two 30-bit full-swing outputs that are generated by this sub-module are complements of each other and drive the ESP1N_M1ARY_PASSGATE_N and ESP1N_M1ARY_PASSGATE_P sub-modules described in the next section.

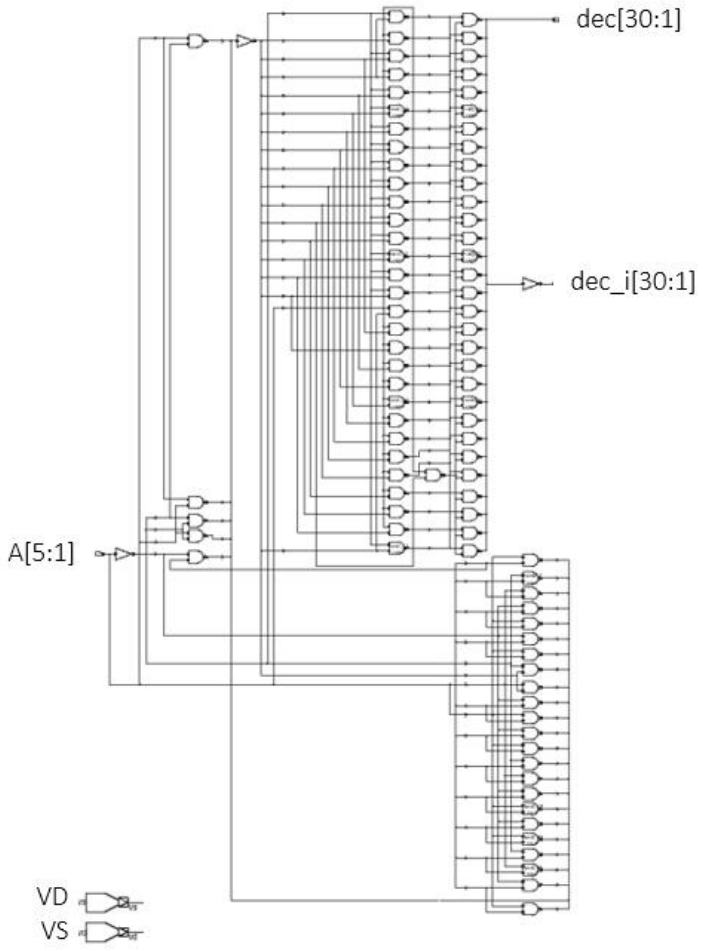

Figure 2.4: ESP1N_M1ARY_DECODE block diagram.

ESP1N_M1ARY_PASSGATE_X

The ESP1N_M1ARY_PASSGATE_N and ESP1N_M1ARY_PASSGATE_P sub-modules are exactly the same and is denoted as ESP1N_M1ARY_PASSGATE_X and depicted in Figure 2.5. As seen in Figure 2.5 the pin list for this sub-module is as follows:

- Two (2) inputs: s and sb.
- Two (2) power supplies: VD and VS.
- Two (2) gate control supplies: VG and VC.
- An output: vout.

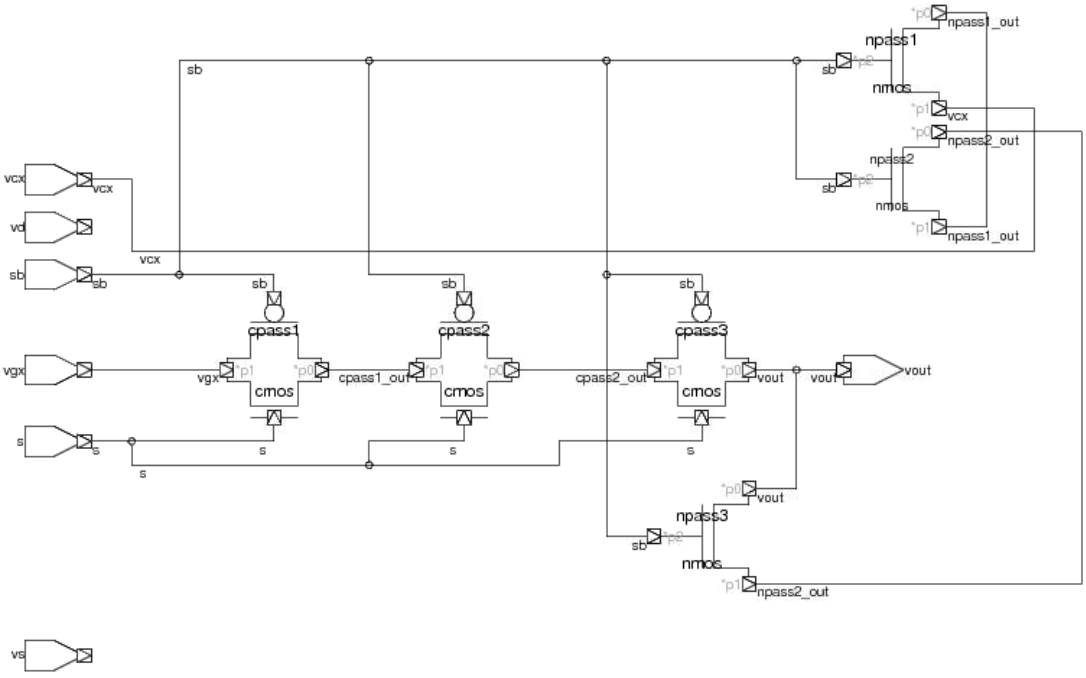

Figure 2.5: ESP1N_M1ARY_PASSGATE_X block diagram.

The ESP1N_M1ARY_PASSGATE_X consists of a series of pass-gates that steers the VG or VC gate control supply to the output vout depending on the two inputs s and sb that are complements of each other and full-swing outputs of the ESP1N_M1ARY_DECODE sub-

module. The VNWELL input in Figure 2.5 is shown as floating but is actually tied to the body connection of the PMOS transistors.

NMOS Arrays

The NMOS_Array_x (where x = 1, 2, 3, 4) sub-module comprises of 30 FinFET NMOS DUTs. A subset of an NMOS_Array_1 sub-module with only four out of the 30 NMOS FinFET DUTs is depicted in Figure 2.6 in the interest of clarity. As seen in Figure 2.6, the pin list for NMOS_Array_1 sub-module is as follows:

- A 30-bit input: g[30:1]
- Two (2) power supplies: VNWELL and VSS.
- An output: drain.

Each of the 30 FinFET NMOS DUTs of an array unit share a common drain and source connection. Each DUT has a unique gate control signal which turns on the DUT and is sourced from either the VGN or VCN supply as per the decoding scheme described above for the ESP1N_M1ARY_A sub-module. The VSS supply powers the source and body terminals of the NMOS DUT, while the VNWELL supply, which is shown as floating in Figure 2.6, is connected to the body of isolated PMOS transistors. The VNWELL connections are not shown in Figure 2.6 but are addressed in Section 2.2.2.

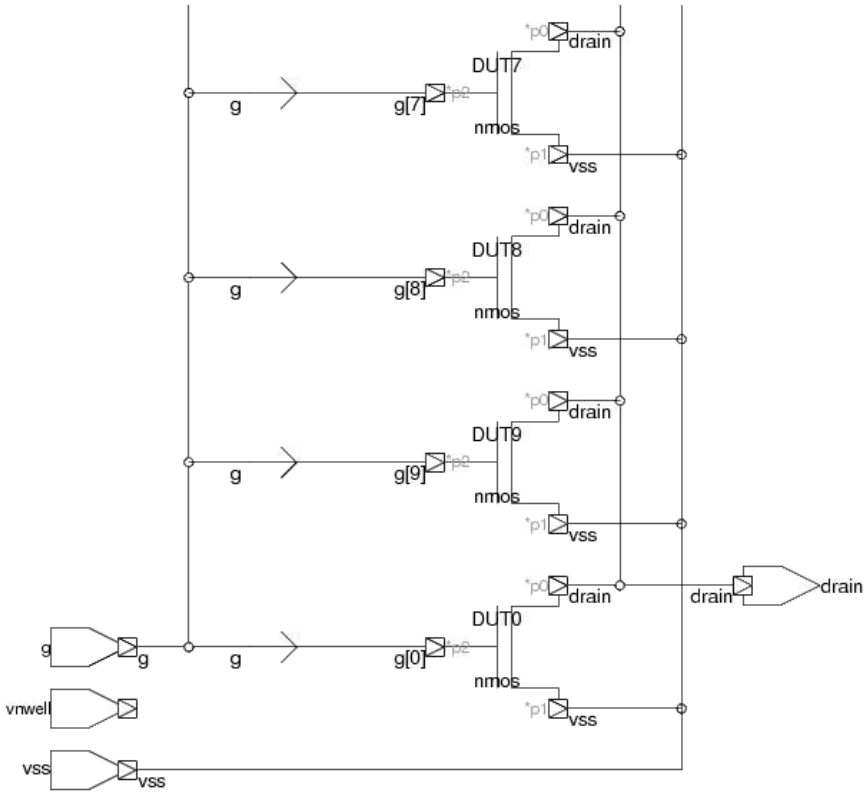

Figure 2.6: Partial NMOS_Array_1 block diagram.

PMOS Arrays

The PMOS_Array_x (where x = 1, 2, 3, 4) sub-module comprises of 30 FinFET PMOS DUTs. A subset of the PMOS_Array_1 sub-module with only four out of the 30 PMOS FinFET DUTs is depicted in Figure 2.7 in the interest of clarity. As seen in Figure 2.7, the pin list for PMOS_Array_1 sub-module is as follows:

- A 30-bit input: g[30:1].
- Three (3) power supplies: VNWELL, VSS and VDD.
- An output: drain.

Each of the 30 FinFET PMOS DUTs of the array unit shares a common drain and source connection. Each DUT has a unique gate control signal which turns on the DUT and is sourced from either the VGP or VCP supply, as per the decoding scheme described above for the ESP1N_M1ARY_A sub-module. The VDD supply powers the source terminal of the PMOS DUT, while the VNWELL supply, which is shown as floating in Figure 2.7, is connected to the body terminal of the PMOS DUT. The VSS supply, which is also shown as floating in Figure 2.7, is connected to the source terminal of isolated NMOS transistors that are not shown in the figure, but are addressed in Section 2.2.2.

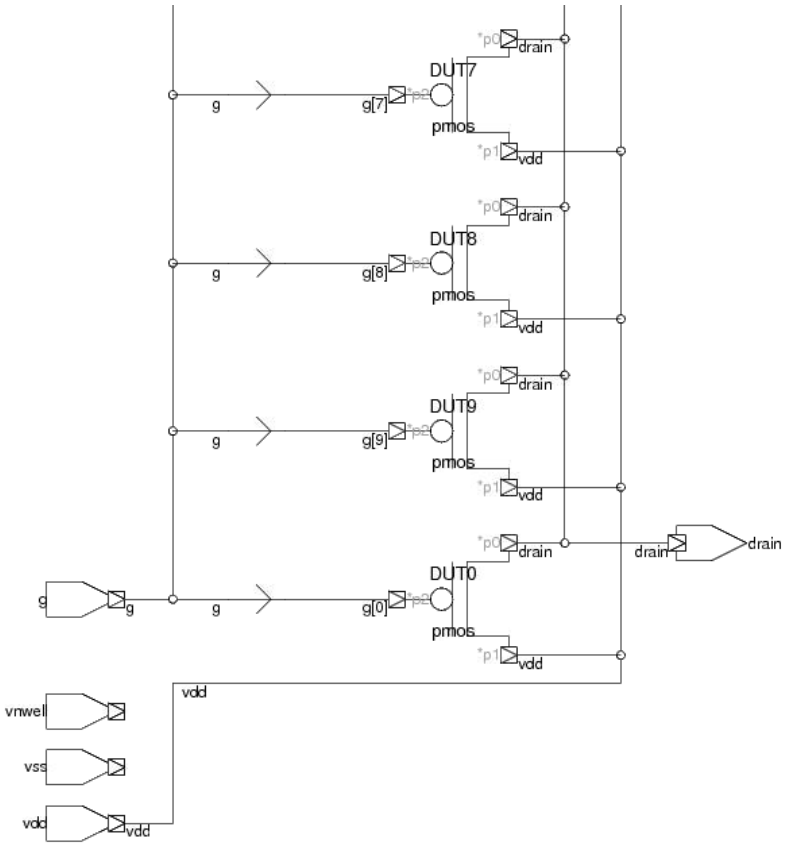

Figure 2.7: PMOS_Array_1 partial block diagram.

2.1.2 DUT Design of Experiment

Formulating a wide-ranging DOE for the DUTs is crucial for a foundry's ability to measure and track test parameters and their variation for a multitude of design-like transistor constructs and layouts. A design-like implementation of a variety of FinFET DUTs that covers attributes like device width, number of fins, V_t types of the DUT and surrounding filler cells, number of fingers, diffusion profiles, and stacking factors will provide the data needed for design modeling and to monitor process variation, yield, manufacturability, and device performance deltas. The SLM test structure presented in this paper features 33 flavors of NMOS and PMOS FinFET DUTs placed across 8 array units. Table 2.1 shows a simplified breakdown of each attribute and number of the DUT DOE. A description for each column of Table 2.1 is as follows:

- DUT #: This is the numerical order of the 33 types of NMOS and PMOS DUTs placed across 8 arrays of the SLM test structure.
- # Fins: This indicates the number of fins of the FinFET transistor in a DUT which can be 2, 4, or 5.
- # Fingers: This is the number of fingers or the number of transistors in parallel for a given DUT that share a source and drain terminal. It can be 1, 2, 4, or stacked. Stacked refers to two transistors in series instead of parallel across a source and drain terminal.
- DUT V_T: This refers to the V_T type of a DUT. It is broadly categorized into Regular V_T (RVT), Low V_T (LVT), and High V_T (HVT).
- Width extend: This refers to an extension of the width of the transistors, which is usually by a nominal amount of about <10% of the original width. A width-extended version of an existing standard cell or transistor is commonly swapped into circuits which may be required to meet stringent leakage or performance

budgets. The SLM test structure features width extended and regular versions of the same DUT.

- Diffusion: This refers to the three types of diffusion profiles a DUT may possess. The *single* type refers to the standard active diffusion profile of a transistor that it shares with neighboring cells, while the *break* type refers to a layout where the active diffusion layer is not continuous but contains a break between neighboring cells. The *L.R break* type refers to a diffusion profile where the left half of the DUT has a wider active diffusion area than the right half. The *R.L break* type is simply the mirror image of the L.R Break diffusion profile.
- Filler V_T: This refers to the V_T type of the transistors in the surrounding or filler cells that abut a DUT and form a barrier to neighboring DUTs. The filler cell V_T for all the DUTs match the DUT V_T except for DUT#2 and DUT#3 in Table 2.1. Section 2.2.2 addresses the filler cell layout of the DUTs in more detail.
- # DUTs: This is a count of the DUTs in both the NMOS and PMOS array units. The total number of DUTs is 120 NMOS and 120 PMOS for a total of 240 FinFET DUTs.

DUT #	# Fins	# Fingers	DUT Vt	Width Extend	Diffusion	Filler Vt	# DUTs
1	4	2	RVT	no	Single	RVT	10
2	4	2	RVT	no	Single	HVT	10
3	4	2	RVT	no	Single	LVT	10
4	4	2	HVT	yes	Single	HVT	3
5	4	2	HVT	no	Single	HVT	3
6	2	2	HVT	yes	Single	HVT	3
7	2	2	HVT	no	Single	HVT	3
8	2	2	LVT	no	Single	LVT	3
9	2	1	LVT	no	Single	LVT	3
10	4	4	RVT	no	Single	RVT	3
11	2	2	RVT	no	Single	RVT	3
12	4	2	RVT	yes	Single	RVT	3
13	2	1	RVT	no	Single	RVT	3
14	4	2	LVT	no	Single	LVT	3
15	2	1	HVT	no	Single	HVT	3
16	4	2	RVT	no	Break	RVT	3
17	2	1	RVT	no	L.R. Break	RVT	3
18	3	1	RVT	no	R.L Break	RVT	3
19	2	stacked	HVT	yes	Single	HVT	3
20	2	stacked	RVT	no	Single	RVT	3
21	2	stacked	RVT	yes	Single	RVT	3
22	2	stacked	HVT	no	Single	HVT	3
23	2	stacked	LVT	no	Single	LVT	3
24	5	2	RVT	no	Single	RVT	3
25	5	2	HVT	no	Single	HVT	3
26	5	2	LVT	no	Single	LVT	3
27	5	2	RVT	yes	Single	RVT	3
28	5	1	RVT	no	Single	RVT	3
29	2	1	RVT	no	L.R. Break	RVT	3
30	4	1	RVT	no	R.L Break	RVT	3
31	5	stacked	LVT	no	Single	LVT	3
32	5	stacked	HVT	no	Single	HVT	3
33	5	stacked	RVT	no	Single	RVT	3

Table 2.1: FinFET DUT design of experiment.

The DUTs were chosen systematically to ensure maximum coverage of commonly used transistor types. The reference cell used to create the DOE is DUT#1 identified as '4Fin, 2Fgr, RVT, noW, Sgl' in shorthand to describe its attributes [i.e., # Fins, # Fingers or stacked, DUT V_T (which typically matches the Filler V_T), width extension (W or noW), and diffusion profile (Sgl, Brk, L.R Brk or R.L Brk)]. Each attribute of the reference DUT

was varied once to come up with a number of experimental splits across all attributes. Table 2.2 details all 32 possible experimental splits created by the DOE along with a shorthand description of the DUTs that comprise each experimental split. Each row of Table 2.2 highlights an experimental split with an 'x' marked in the columns of the DUT# to show the DUTS in that split. Each DUT can be part of multiple experimental splits in the SLM structure. The DOE offers the ability to compare the same attributes for multiple flavors of the FinFET transistor allowing multiple-way comparison of key test parameters across the entire standard-cell design space. This DOE implementation thus allows for the monitoring of a wide range of true design-like implementations of FinFET transistors on the SLM test structure.

Attribute	Experimental split	DUT #																															DUT attributes for experiment splits
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	
# Fins	2 vs. 4	x										x																							2Fgr, RVT, noW, Sgl
	2 vs. 4				x		x																												2Fgr, HVT, W, Sgl
	2 vs. 4					x		x																											2Fgr, HVT, noW, Sgl
	2 vs. 4								x						x																				2Fgr, LVT, noW, Sgl
	1 vs. 2																								x				x						5Fgr, RVT, noW, Sgl
# Fingers	2 vs. 4	x									x																								4Fin, RVT, noW, Sgl
	1 vs. 2							x								x																			2Fin, HVT, noW, Sgl
	1 vs. 2								x	x																									2Fin, LVT, noW, Sgl
	1 vs. 2											x		x																					2Fin, RVT, noW, Sgl
	1 vs. 2																								x				x						5Fin, RVT, noW, Sgl
Stacked	Y vs. N						x													x															2Fin, HVT, W, Sgl
	Y vs. N							x															x												2Fin, HVT, noW, Sgl
	Y vs. N											x									x														2Fin, RVT, noW, Sgl
	Y vs. N																								x									x	5Fin, RVT, noW, Sgl
	Y vs. N																									x							x		5Fin, HVT, noW, Sgl
	Y vs. N																										x					x			5Fin, LVT, noW, Sgl
Diffusion	Sgl vs. Brk	x															x																		4Fin, 2Fgr, RVT, noW
	L.R vs. R.L																	x	x																2-3Fin, 2Fgr, RVT, noW
	L.R vs. R.L																													x	x				2-4Fin, 1Fgr, RVT, Sgl
Width Extend	Y vs. N	x											x																						4Fin, 2Fgr, RVT, Sgl
	Y vs. N				x	x																													4Fin, 2Fgr, HVT, Sgl
	Y vs. N						x	x																											2Fin, 2Fgr, HVT, Sgl
	Y vs. N																			x			x												2Fin, stacked, HVT, Sgl
	Y vs. N																				x	x													2Fin, stacked, RVT, Sgl
	Y vs. N																								x			x							5Fin, 2Fgr, RVT, Sgl
DUT Vt	L vs. R vs. H	x				x									x																				4Fin, 2Fgr, noW, Sgl
	L vs. R vs. H							x	x			x																							2Fin, 2Fgr, noW, Sgl
	L vs. R vs. H									x				x		x																			2Fin, 1Fgr, noW, Sgl
	L vs. R vs. H																				x		x	x											2Fin, stacked, noW, Sgl
	L vs. R vs. H																								x	x	x								5Fin, 2Fgr, noW, Sgl
	L vs. R vs. H																															x	x	x	5Fin, stacked, noW, Sgl
Filler Vt	L vs. R vs. H	x	x	x																															4Fin, 2Fgr, RVT, noW, Sgl

Table 2.2: FinFET DUT experimental splits.

2.2 DESIGN AND IMPLEMENTATION

The design flow of the SLM test structure was comprised of Verilog modeling followed by circuit design and layout work. Once the design work was completed, the full netlist was extracted to conduct extensive HSPICE simulation of the DUTs for all test parameters. The results of the simulation and test plan for the SLM test structure is covered in Chapter 3. This section goes over the results of the Verilog validation work, followed by a discussion of the various unique features of the circuit and layout design process for the array units and DUTs.

2.2.1 Verilog Validation

The SLM test structure was modeled in Verilog as per the design of the sub-modules detailed in Section 2.1. Since the design was made purely combinational to allow rapid DC I-V measurements of the DUTs the Verilog model contains no sequential elements like latches, flops, or state machines. Figure 2.8 depicts the hierarchy of the AMDFETH1A module in the test bench for the purposes of Verilog modeling. The ESP1N_M1ARY_DECODE_X sub-module has 30 underlying ESP1N_M1ARY_PASSGATE_X sub-modules not shown in Figure 2.8. The test bench for the Verilog validation step was coded as part of the module named top which also instantiates the AMDFETH1A module. The test bench consists of a regression of tests to verify the correct operation of the 5-bit decoder for all input combinations and to verify the power and gate connections to the DUTs for all the array units to ensure proper functionality.

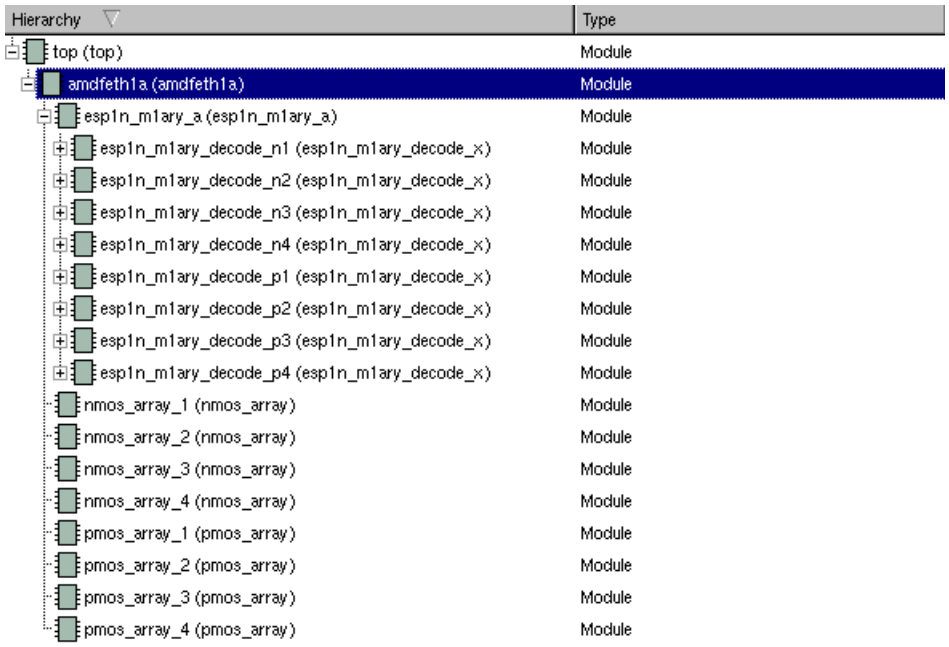

Figure 2.8: Verilog AMDFETH1A model hierarchy.

Figure 2.9 shows the waveforms from an example Verilog validation test of the AMDFETH1A design. Figure 2.9 also depicts the standard operation of the SLM test structure. Details on the test plan and operation of the SLM test structure are also covered in Chapter 3 of this paper. As shown in Figure 2.9, all supplies to the array units (VDD_N1TO4, VDD_P12, and VDD_P34) are initially low and all gate control supplies (VGN, VCN, VGP, and VCP) are set to turn off the gate of the DUTs. The drain supplies DR_Nx (where x = 1, 2, 3, 4) are modeled purely as outputs for Verilog validation. To initiate the selection of a DUT, the decoder input A[5:1] (shown as a[4:0] in Figure 2.9) is set to b'00100 (decimal 4), followed by the VDD_N1TO4 supply going high to select the NMOS array units. At the 3000ns time period, VGN is made to go high to turn on the gate to the selected DUT. The result of the VGN signal going high is seen at the 4000ns time period in Figure 2.9 where the N1_sel[30:1] (shown as n1[29:0] in Figure 2.9) DUT gate

selection signal settles to h'00000008 effectively steering the VGN signal to the selected DUT. At this point all four selected DUTs in the NMOS arrays 1 through 4 will be turned on and all the drain outputs (DR_N1, DR_N2, DR_N3, and DR_N4) settle to a low indicating that current is flowing through the source and drain of the selected DUTs.

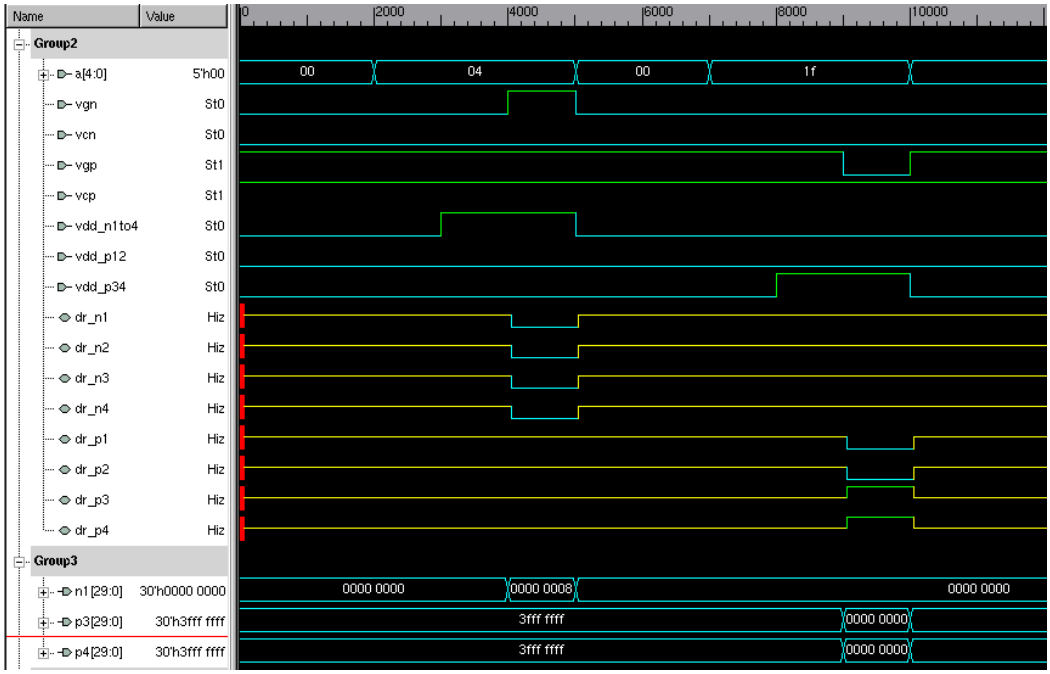

Figure 2.9: Verilog AMDFETH1A validation example.

In the latter half of this test, the VDD_N1TO4 and VGN supplies are turned off followed by the decoder input A[5:1] (shown as a[4:0] in Figure 2.9) being set to h'1F (decimal 31) to select all 30 DUTs of an array unit. VDD_P34 supply goes high at time period 8000ns and then 1000ns later, the VGP supply goes low to turn on the selected DUTs of PMOS array units 3 and 4. At this point the P3_sel[30:1] (shown as p3[29:0] in Figure 2.9) and P4_sel[30:1] (shown as p4[29:0] in Figure 2.9) signals settles to h'00000000 to signify the steering of VGP signal to the gate terminals of all 30 PMOS DUTs of PMOS array 3 and

4. The DR_P3 and DR_P4 outputs are driven high as a result of the gates of all the PMOS DUTs being turned on. At the same time, DR_P1 and DR_P2 also go low since, although the gates of all the PMOS DUTs in PMOS arrays 1 and 2 are also being selected, the power supply VDD_P12 to these two arrays are turned off and the DUTs are drawing no current from the source to the drain. The DR_P1 and DR_P2 outputs going low is merely an artifact of the Verilog model and won't be seen in the actual design implementation. The Verilog validation of the design was an important step to prove the decoder circuit functionality and to help define the SLM test structure operation and the overall test plan. It also enabled the generation of a set of test vectors to verify the final netlist extract of the design after layout.

2.2.2 Circuit Design and Layout

The circuit design and layout for the SLM test structure were two parallel efforts. The decoder circuits were essentially ported from an earlier design and synthesized for the new 14nm FinFET technology, while the 8 NMOS and PMOS array units were handcrafted using existing standard cells from the 14nm FinFET technology library. The decoders and the array units were then integrated to complete the design. This section starts with the details of the overall floorplan followed by a discussion of the design and layouts of the array unit, DUT, and filler cells.

SLM Test Structure Floorplan

The SLM test structure was floorplanned to fit an existing 1x25 pad SLM design configuration already in use at the foundry with an existing inline test infrastructure consisting of probe cards and test instruments designed to drive and measure signals to

specific pads of the design. The floorplan for the 52µm by 2000µm SLM test structure is depicted in Figure 2.10. The layout consists of 25 input-output pads, 8 array units and their decoder circuitry, 5 electrostatic discharge (ESD) cells for the decoder input A[5:1], and a bus of decoder signals, gate voltage control supplies, power, and ground signals that run along the length of the design. The drain and source pads connected to the array units are hooked up to the power and ground supplies and also have a level of ESD protection to ensure electrical integrity while still maintaining enough fidelity to conduct current measurements on a transistor scale.

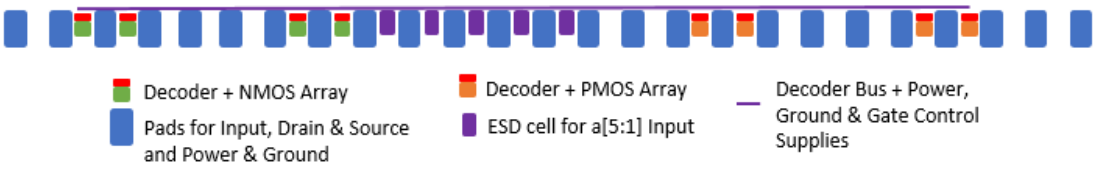

Figure 2.10: SLM test structure floor-plan (not to scale).

The decoder inputs and the NMOS and PMOS array units were placed next to pads that were already designated by the foundry test specifications as inputs and outputs. The power, ground, body, and gate control supply signals are also connected to pre-designated pads of the SLM test structure. The decoder circuitry for each array unit, which consists of the ESP1N_M1ARY_DECODE_X module, was ported and synthesized to fit in the area between the pads with enough room for the array units. For the decoder circuitry, as described in Section 2.1.1, the ESP1N_M1ARY_DECODE and ESP1N_M1ARY_PASSGATE layout was crafted such that the pitch of the 30-bit output signal matched the pitch of the gate control lines of the array unit. Special care was also taken to design the pass gates of the ESP1N_M1ARY_PASSGATE sub-module using low leakage transistors to steer gate control voltages with minimal voltage drop on the lines.

While the decoder circuitry was a straight-forward port of existing designs, the array unit layout was done from scratch and is discussed below.

Array Unit Layout

The array units of the SLM test structure were designed to fit 30 DUTs in a 3x10 matrix with filler cells surrounding each DUT and filler column cells between each column. Figure 2.11 depicts a simplified representation of the layout of an array unit along with the decoder circuitry and the power supplies.

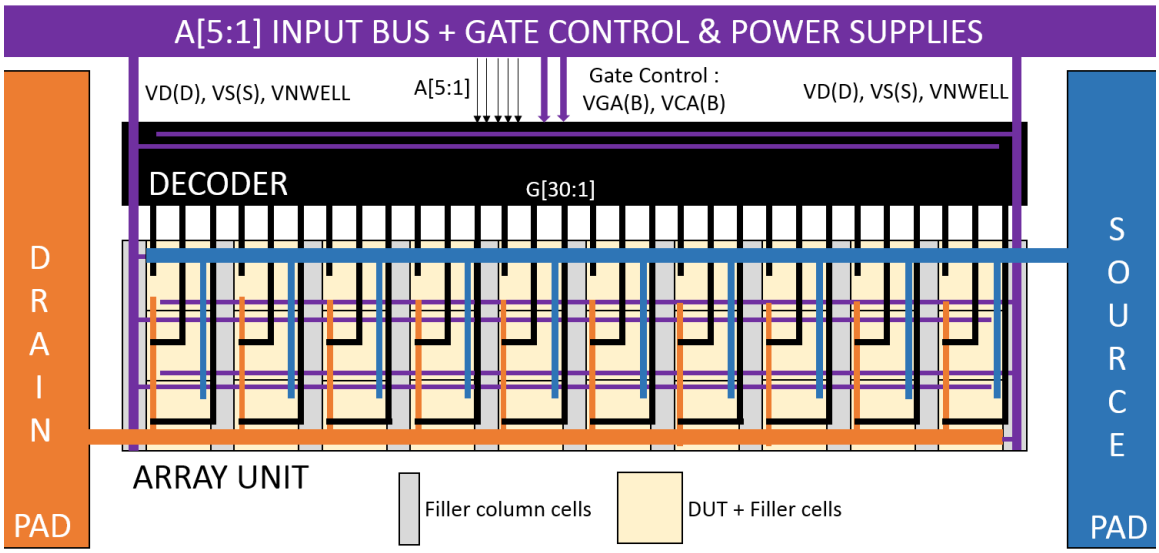

Figure 2.11: Array unit and decoder layout representation (not to scale).

As seen in Figure 2.11, the array unit comprises of 30 DUTs with filler cells and 11 filler cell columns between the DUT columns. Each DUT consists of one of the types of DUT flavors listed in Table 2.1. An array unit typically has multiple copies of the same type of DUT usually either 3 or 10 in number. All 30 DUTs of the array unit are connected to the

same source and drain signals which run horizontally across the array units. Although not depicted in Figure 2.11, the drain and source pad connections to the DUTs comprise a grid of M2 and M4 lines designed to minimize contact resistance at the pad connections during measurement. The source and drain connections for all DUTs are connected separately via an M2 grid within the array unit and then tapped into wider M4 grids that occupy distinct portions of the array unit and ultimately connect to the pads. This is represented in Figure 2.11 by the orange and blue vertical lines for the M2 grid, and horizontal lines for the M4 grid over the array unit.

The decoder circuitry for the array units was designed to take up roughly 30% of the area of the array unit but was made wide enough to line up the gate connection outputs (G[30:1]) to the gate inputs for each DUT depicted by the black vertical lines into the DUTs in Figure 2.11. The G[30:1] connections run in vertical M3 lines. The gate control voltage supplies (VGA and VCA or VGB and VCB depending on the NMOS or PMOS transistor type), decoder inputs (A[5:1]), and the various power, ground, and body supplies (VDD, VD, VSS, VS, and VNWELL) to the decoder circuit and the array units run along the length of the SLM test structure above the pads and tap into each decoder and array unit. Although not shown in Figure 2.11, the power, ground, and body supplies utilize separate power planes of either M2, M3, or M4 which run orthogonally to each other but always in a single orientation. These lines occupy distinct portions of the layout to tap into the requisite cells of the decoder and array unit. Such a design helps mitigate potential issues with capacitive loading or voltage droops. The connections for these supplies is represented by the purple lines in Figure 2.11. For all power, source, and drain lines, the widths of the metal lines and via connections were made as wide as possible within design rule checker (DRC) constraints. Any unnecessary jogs or 90-degree angles in metal lines for the array unit and its connections to the decoder circuit were avoided. This was made possible by

careful selection of the filler cell and DUT widths to ensure a uniform pitch for the G[30:1] M3 lines. The dimensions and layout of the array unit and the decoder circuitry were ultimately determined by the layout of the DUT and filler cells.

The NMOS and PMOS DUTs were implemented using standard cell inverters with isolated transistors. Figure 2.12 shows a schematic view of the PMOS and NMOS DUT in the array units of the SLM test structure.

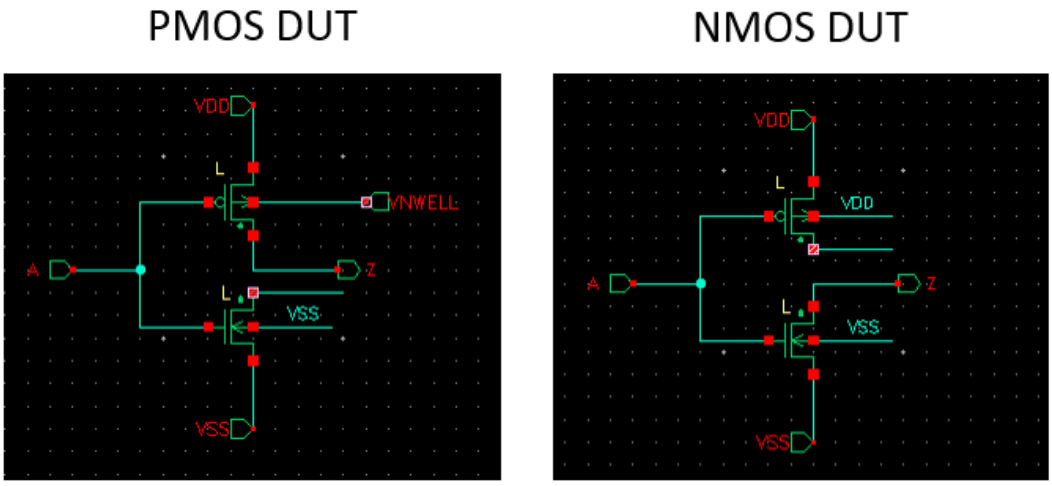

Figure 2.12: PMOS and NMOS DUT schematics.

The PMOS DUT consists of an isolated NMOS transistor, while the PMOS transistor remains connected to the output terminal of the inverter. The body of the NMOS is tied to VSS, while the body of the PMOS is tied to VNWELL. The NMOS DUT consists of an isolated PMOS transistor, while the NMOS transistor remains connected to the output terminal of the inverter. The body of the NMOS is tied to VSS, while the body of the PMOS is tied to VDD. Figure 2.13 shows an example of an NMOS DUT layout without any filler

cells and with labeled terminals. To isolate the PMOS transistor in the inverter the Via0 contact to M1 layer was removed. Such an implementation of a design-like standard cell will allow the SLM structure to produce the most relevant transistor parametric data for the foundry.

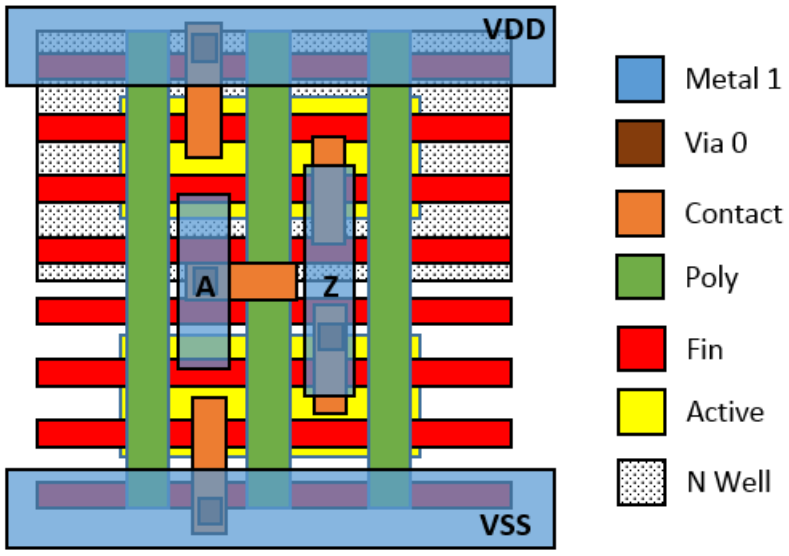

Figure 2.13: NMOS DUT without filler cells.

An example layout of the NMOS DUT with the filler cells is shown in Figure 2.14. The filler cells in the DUTs and the column cells are electrically isolated even though they are abutted to the inverter, share the same VDD and VSS M1 lines, and share the N-Well diffusion and active layers. The filler cells are essentially stacked untapped transistors abutted to the standard cell used to implement the DUT. There are no vias or metal lines over the filler cells of the DUT or the columns. This allows for easy routing of the gate control lines and the source and drain grids over the DUTs. Only the M1 VDD and VSS lines of the filler cells that are abutted to the DUT are powered since they share the same

metal lines. These M1 VDD and VSS lines of the DUT are connected to the VDD and VSS power lines of the bus through a grid of M2, M3, and M4 lines. The remaining M1 VDD and VSS lines of the top and bottom rows of the filler cells surrounding the DUT are unconnected to these grids. The filler column cells in Figure 2.11 follow the same layout style and constructs as the filler cells that surround the DUT.

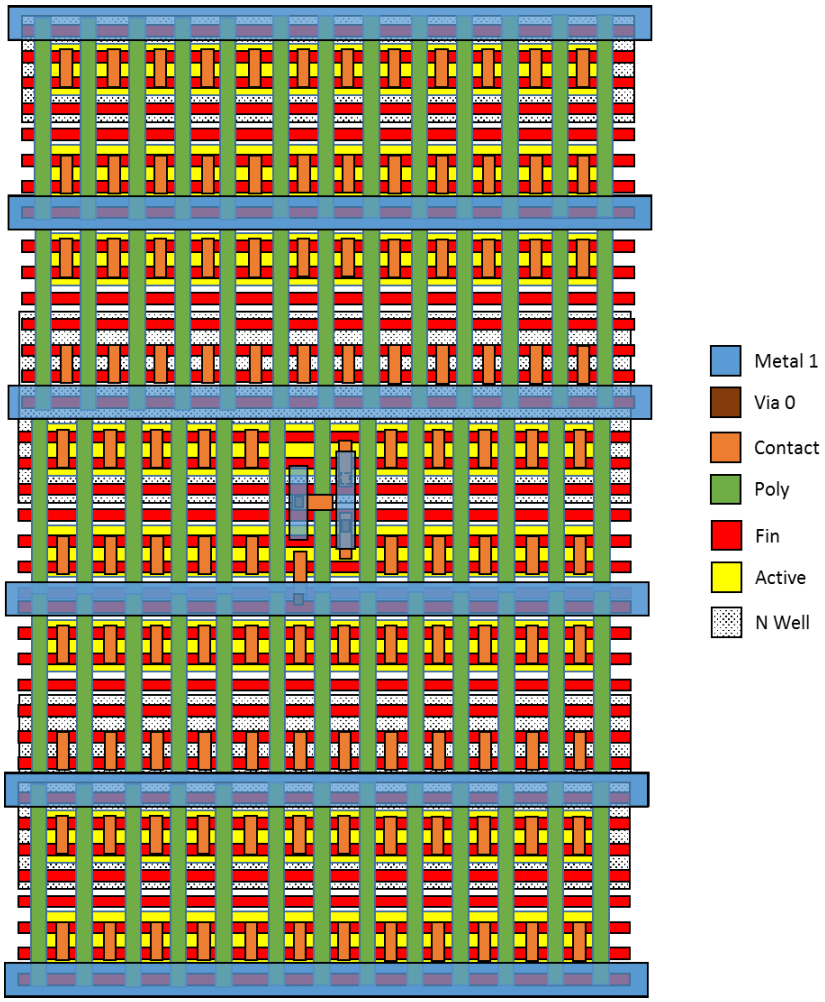

Figure 2.14: NMOS DUT with filler cells.

An important metric during the layout of the array units was the poly width counts of the DUT and filler cells. Optimizing the poly width count of each array unit, which contained multiple DUTs and filler cells of varying widths, allowed the array unit to be sized as uniformly as possible to ensure the gate connections from the decoder circuitry always lined up with the DUTs. This negated the need to have a separate layout for each decoder circuit and saved considerable effort. A major advantage of filler cells was that it permitted enough room for the gate control wiring for all DUTs to be done uniformly in a single metal layer. It also allowed the DUT to be self-contained and isolated from other DUTs with no need for additional higher level metal local interconnects or poly jogs. The only layout modification needed to implement the NMOS or PMOS transistor in the DUT was the removal of a single via, thereby retaining the original design-like standard cell architecture inline measurement. The implementation of the DUT with filler cells allows secondary but real effects like stress from neighboring cells and the variation due to the non-uniformity of the layout to be reflected in the transistor measurements. These features will enable the collection of FinFET transistor parameters that are expected to match the actual design and thus offer the best prediction for chip yield, transistor design models and process variability.

2.2.3 Post-Layout

Once all unit arrays, decoder circuitries, ESD cells, and input and power buses were integrated, the SLM test structure design was ready for the required post-layout steps. DRC and layout versus schematic (LVS) checks were at done at every hierarchical stage during the layout phase and after the full design was integrated. A list of known DRC violations is to be catalogued for the foundry to review. Most of these violations were expected due

to the unconventional implementation of the standard cells and filler cells in the array units typically with respect to boundary edges of the array unit. It is expected that the foundry will either waive these violations for the SLM test structure or require further changes to the layout. Following a clean LVS and final DRC check, special front-end-of-line and back-end-of-line fill algorithms were run to selectively add metal, active, and poly layers to meet certain uniform density gradient metrics to meet DFM rules of the foundry.

After a final Verilog versus layout test bench validation of the design, the next step is to complete the netlist extraction for HSPICE simulation of all test parameters. At this point, the design layout database is ready to be delivered to the foundry. In addition to the database, the foundry also expects a comprehensive test plan to cover the testing methodology, specifications for all supplies and inputs, and a list of all the parameters to be measured. The HSPICE simulation results and a detailed inline test plan are presented in the following chapter.

Chapter 3: Inline Test Plan and Simulation Results

This chapter details the inline test plan for the SLM test structure and the HSPICE simulation results for key test parameters for a few DUT types. One of the goals of the SLM test structure is to evaluate actual silicon data versus simulation data to improve the technology library and design models. This will also help validate the test infrastructure and the design. The architecture and combinational nature of the SLM test structure lends itself to rapid and parallel DC testing for the test parameters that are addressed in detail in the following section. A discussion of the test algorithms and various aspects of the inline test plan are also presented in the following section.

3.1 Test Parameters and Inline Test Plan

The test parameters for the SLM test structure are limited to I-V parameters like source-drain current and threshold voltage measurements for all DUT types. These measurements will also allow for the calculation of secondary parameters that can be post-processed on the tester. A listing of test parameters for each NMOS DUT that is to be normalized to a 'number of fins * number of fingers' metric under a range of test voltage conditions (Vdd) is as follows:

- I_{DSAT}: Source-drain current in the saturation region of the transistor with both source and gate terminals at the Vdd value.
- I_{DLIN}: Source-drain current in the linear region of the transistor with the gate terminal set to Vdd and the drain terminal at a fixed 5% of the maximum Vdd value.
- I_{DEFF-H}: Source-drain current in the linear region of the transistor with the gate terminal set to Vdd and the drain terminal at a fixed 50% of the maximum Vdd value.

- I_{DEFF-L}: Source-drain current in the linear region of the transistor with the drain terminal set to Vdd and the gate terminal at a fixed 50% of the maximum Vdd value.
- I_{DEFF}: Average of the previous two parameters I_{DEFF-H} and I_{DEFF-L}.
- I_{DOFF}: Source-drain current when the transistor is turned off with gate terminal at ground and drain terminal set to Vdd.
- V_{TSAT}: Threshold voltage in the saturation region of the transistor at a pre-defined technology dependent source-drain current value for a given number of fins and number of fingers for the DUT. The drain terminal is set to Vdd as the gate is swept until the desired current is observed.
- V_{TLIN}: Threshold voltage in the linear region of the transistor at a pre-defined technology dependent source-drain current value for a given number of fins and number of fingers for the DUT (this is the same value as the current level for the V_{TSAT} parameter). The drain terminal is set to a fixed 5% of the maximum Vdd as the gate is swept until the desired current is observed.

The source terminal for the NMOS DUT is always set to ground. The parameters for the PMOS DUT are analogous to the NMOS DUTs. The measurement algorithms for the test parameters are prescribed for predefined voltage supply (Vdd) settings that are applied to the specified input supplies of the SLM test structure. These terminals are the gate voltage supplies VGA and VCA for NMOS array, VGB and VCB for the PMOS array, and drain control supplies DR_Nx and DR_Px (where x = 1, 2, 3, 4).

3.1.1 Measurement Algorithms

An example measurement algorithm with hypothetical voltage values for NMOS array 1 is given by the following pseudo-code under the assumption that the source-drain current

for the selected DUT with the transistor turned on is about an order of magnitude higher than the total leakages of all the unselected DUTs in the array unit.

- For each Vdd setting from 500mV to 1V in 200mV steps:
    1. For each selected single DUT using A[5:1] input:
        i. Set VCA to -200mV to turn off unselected DUT.
        ii. Set DR_N1 to 50mV and sweep VGA from 0V to current Vdd and measure $I_{DS.}$ Record V_{TLIN} as the VGA value that satisfies the predefined I_{DS} condition as described for the V_{TLIN} test parameter earlier.
        iii. Set DR_N1 to 50mV and VGA to Vdd. Measure current sourced through DR_N1 to determine I_{DLIN}.
        iv. Set DR_N1 to 500mV and VGA to Vdd. Measure current sourced through DR_N1 supply to determine I_{DEFF-H}.
        v. Set DR_N1 and VGA to Vdd and measure current sourced through DR_N1 supply to determine I_{DSAT}.
        vi. Set DR_N1 to Vdd and VGA to 500mV. Measure current sourced through DR_N1 supply to determine I_{DEFF-L}.
        vii. Set DR_N1 to Vdd and sweep VGA from 0V to current Vdd and measure $I_{DS.}$ Record V_{TSAT} as the VGA value that satisfies the predefined I_{DS} condition as described for V_{TSAT} test parameter earlier.
        viii. Set DR_N1 to 50mV and sweep VGA from 0V to current Vdd and measure $I_{DS.}$ Record V_{TLIN} as the VGA value that satisfies the predefined I_{DS} condition as described for the V_{TLIN} test parameter earlier.

ix. Calculate $I_{DEFF} = (I_{DEFF\text{-}H} + I_{DEFF\text{-}L)} / 2$.

2. Select all DUTs in the array unit using A[5:1] input = b'11111.
    i. Set VGA gate voltage supply to 0V and VCA to -200mV.
    ii. Set DR_N1 to Vdd and measure $I_{DOFF\text{-}ALL}$ as the current sourced though DR_N1.
3. Select no DUTs in the array unit using A[5:1] input = b'00000.
    i. Set VGA gate voltage supply to 0V and VCA to -200mV.
    ii. Set DR_N1 to Vdd and measure $I_{DOFF\text{-}NONE}$ as the current sourced though DR_N1.
4. For each selected DUT in the array unit using A[5:1] input:
    i. Set VGA gate voltage supply to 0V and VCA to -200mV.
    ii. Set DR_N1 to Vdd and measure $I_{DOFF\text{-}MIXED}$ as the current sourced though DR_N1.
    iii. Calculate $I_{DOFF} = (I_{DOFF\text{-}MIXED} - I_{DOFF\text{-}NONE}) / (1 - (I_{DOFF\text{-}NONE} / I_{DOFF\text{-}ALL}))$.

For array units, where simulation of silicon data suggests that the total leakage of the unselected DUTs in the array unit is closer to the on-current of the selected DUT, a minor correction will need to be performed as described in the original SLM FET array design in Section 1.2 [9]. It is important to note that the I_{DOFF} measurements can only be performed on array units that contain a single type of DUT.

3.1.2 Inline Test Plan

The inline test plan for the foundry will comprise of the measurement algorithm described above but for all array units along with a catalog of the test conditions,

architecture information to determine pad dimensions and location for all supplies, inputs and outputs, desired format of the data-logs, site information of the SLM test structure on the wafer, and other yet to be determined information for the data-logs. The target tester for the foundry will be an Agilent 4082 parametric test system that is designed to conduct rapid and precise DC measurements. For DC measurement purposes, the tester contains multiple self-calibrating, ultra-low current instruments capable of measurements in the 1fA to 100mA range with an accuracy in the 1% range or lower depending on the precision settings as a trade-off for test time. The tester also has built-in circuitry to perform voltage sweeps and concurrently measure current on voltage pins in parallel at multiple test insertions in the manufacturing line. The pins in the probe cards for the test system may need to be modified depending on their compatibility with the probe pad dimensions of the SLM for the 14nm FinFET technology library. The probe card pin configuration is expected to be compatible for 1x25 pad pin placement by design. Depending on the number of SLM test structures placed on each wafer (expected to be between 10 and a 100), the test algorithms will most likely be configured to run all 8 array units in parallel, except in the case of I_{DOFF} test parameter measurement, to speed up testing.

The testing will occur at M4 and at room temperature although the foundry may conduct hot or cold testing as a special case. The test program code for the Agilent 4082 will be directly based on the algorithms provided and will be written by assigned test engineers at the foundry. The test program will generate data-logs that will contain information about the test site, test conditions, array number, and the DUT type populated by the tester code. Measurement values populated from the tester instruments and calculated values based on the measurement values will also be a part of the data-logs. The test data-log for each wafer is expected to contain at least the following information:

- Header

    - Lot Number
    - Product ID
    - X coordinate
    - Y coordinate
    - SITE location reference number on the wafer
    - Scribe ID number for the SLM test structure
    - Temperature
- Data for each DUT
    - Array number under test
    - Array type under test
    - Vdd test voltage value
    - Array $I_{DOFF\text{-}MIXED}$ measurement value (if applicable)
    - Array $I_{DOFF\text{-}NONE}$ measurement value (if applicable)
    - Array $I_{DOFF\text{-}ALL}$ measurement value (if applicable)
    - DUT number under test
    - DUT shorthand name under test
        - Number of fins
        - Number of fingers
        - V_T type
        - Width extension case
        - Diffusion case
        - Length
        - Width
    - DUT I_{DSAT} measurement value
    - DUT I_{DLIN} measurement value

- DUT I_{DSAT} measurement value
- DUT I_{DEFF-H} measurement value
- DUT I_{DEFF-L} measurement value
- DUT I_{DEFF} calculated value
- DUT I_{DOFF} calculated value (if applicable)
- DUT V_{TSAT} measurement value
- DUT V_{TLIN} measurement value

3.2 Simulation Results

One of the main goals of the SLM test structure is to bridge any gaps between the current technology models and actual silicon data. Full HSPICE simulation runs are underway on the SLM test structure to establish a baseline for all DUTs across all parameters and all corners of the process. The simulation results help assess the relative differences in the test parameters between all the DUT types and also helps to validate the design and the test algorithms. Upon completion of the design, the layout and schematic netlist for all array units were extracted using the latest technology library models. HSPICE decks from previous projects were leveraged to run simulations on the layout and schematic extracts of the 14nm FinFET SLM design. Certain 14nm FinFET related updates to the HSPICE decks were necessary such as voltage, temperature, and new formulae for V_{TSAT} and V_{TLIN} calculations. All HSPICE simulations were conducted across multiple process corners: TT (typical), SS (slow) and FF (fast). Some key normalized simulation results for the NMOS DUT across all the experiments of the DOE are presented in the following sections.

3.2.1 Current and Threshold Voltage Trends

Figure 3.1 shows the results of the simulation of NMOS DUT#1 (4Fin, 2Fgr, RVT, noW, Sgl) from Table 2.1 for the TT process corner for the various source-drain current parameters normalized to a 0 to 1 scale. The bottom axis is the Vdd test voltage setting, which corresponds to the terminal voltage as described in the measurement algorithm in the previous section, and the left axis are the normalized current values. As seen in Figure 3.1, I_{DSAT} and I_{DEFF} scales with the Vdd drain voltage setting as expected, while the I_{DLIN} and I_{DOFF} shows a comparatively flatter response. I_{DLIN} saturates at higher Vdd, while I_{DOFF} values increase rapidly at the higher end of the drain voltage setting.

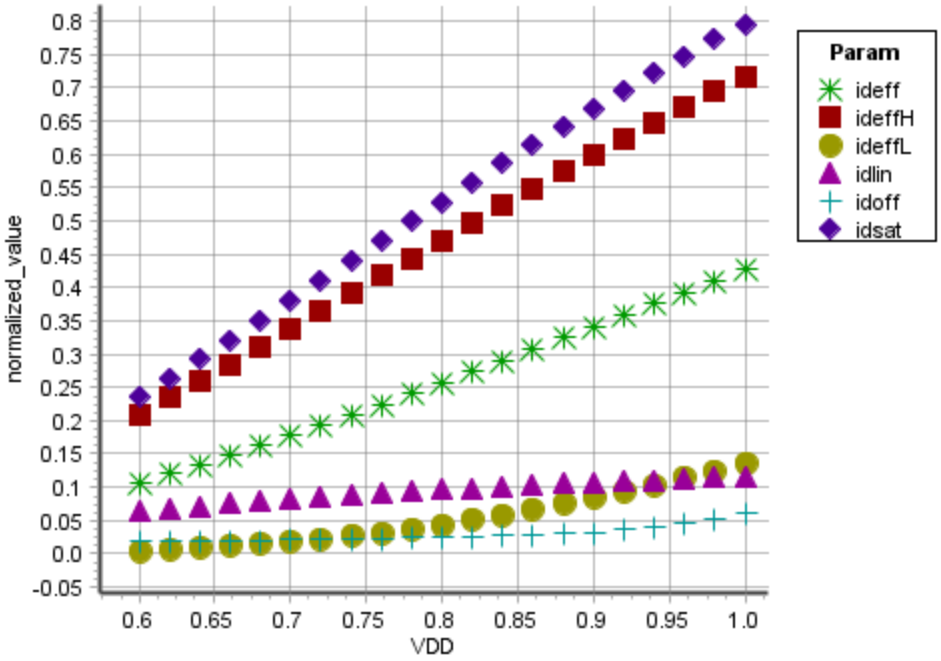

Figure 3.1: NMOS DUT#1 normalized current simulation results.

Figure 3.2 shows the normalized results of the same NMOS DUT#1 but for the threshold voltage parameters in the TT process corner. The V_{TLIN} parameter is higher than V_{TSAT} as expected since the drain voltage is set to a fixed 5% of the maximum Vdd setting as the gate voltage is swept until the I_{DS} condition for this DUT is met. It takes a higher gate voltage in the linear region to meet the same I_{DS} as in the saturation region hence the V_{TLIN} is higher. The V_{TLIN} trend with Vdd setting is irrelevant since the drain setting is a fixed value of 5% of the maximum Vdd setting. The V_{TSAT} shows a slight inverse linear trend versus the drain voltage setting Vdd suggesting that the higher the drain-source voltage, the higher the drain-source current to meet the I_{DS} condition.

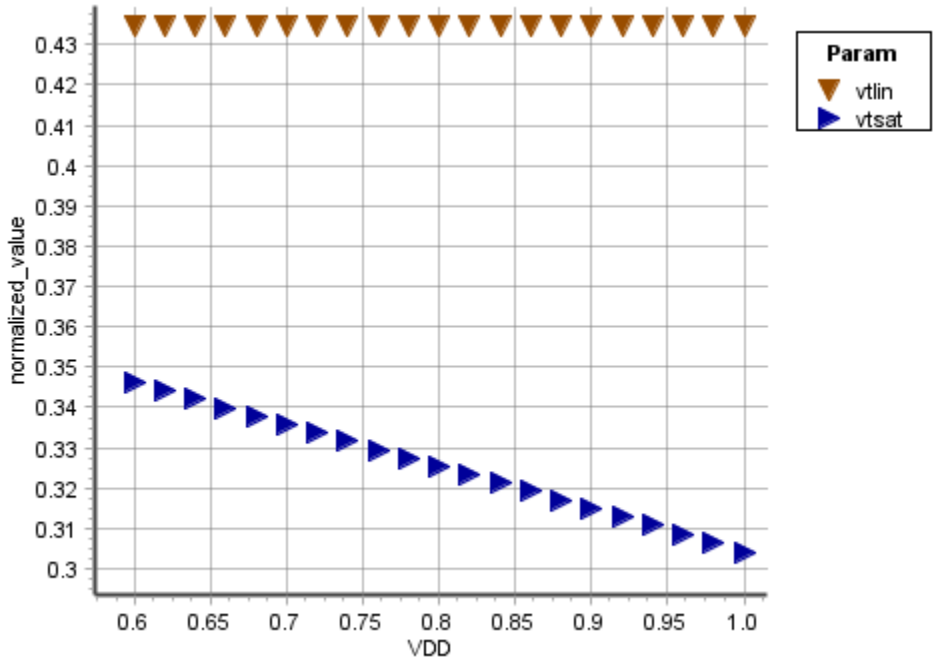

Figure 3.2: NMOS DUT#1 normalized threshold voltage simulation results.

3.2.2 Key DUT Simulation Trends

Figure 3.3 shows a comparison of two NMOS DUTs with a different number of fins for the TT process corner. NMOS DUT#1 (4Fin, 2Fgr, RVT, noW, Sgl) and DUT#11 (2Fin, 2Fgr, RVT, noW, Sgl) from Table 2.1 were simulated to show the difference in test parameters for the number of fins attribute.

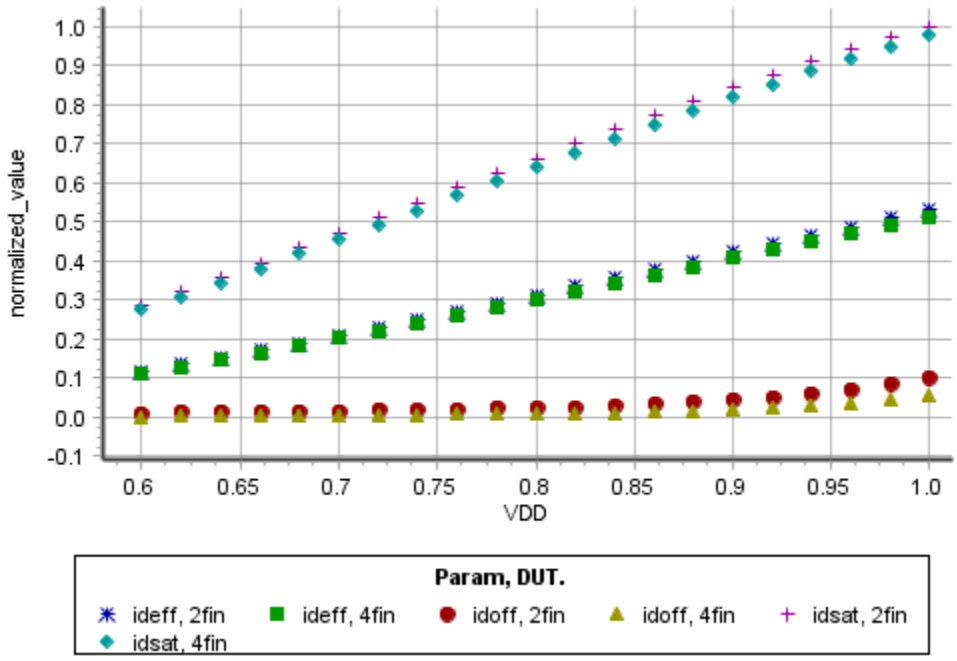

Figure 3.3: NMOS number of fins normalized current simulation results.

In Figure 3.3, DUT#11 shows nearly equal normalized values for I_{DSAT}, I_{DEFF}, and I_{DOFF} as DUT#1. In general, we expect higher drive with higher number of fins, but this wasn't borne out in the simulation results for these DUTs and remains to be seen in the silicon data. The layout and library models will be the subject of further investigation as we try to understand the expected trends and values with respect to the number of fins. A similar

simulation result was observed for the number of fingers and the width extension versus no width extension experiments. These trends will need to be investigated as well. Figure 3.4 shows a comparison for the three V_T types: HVT, LVT, and RVT. NMOS DUT#12 (2Fin, 2Fgr, RVT, noW, Sgl), DUT#8 (2Fin, 2Fgr, HVT, noW, Sgl), and DUT#7 (2Fin, 2Fgr, LVT, noW, Sgl) from Table 2.1 were simulated to show the variation in the voltage threshold test parameters across V_T types. As seen in Figure 3.4, the HVT DUT has higher normalized V_{TSAT} and V_{TLIN} values than the RVT DUT followed by the LVT DUTs as expected.

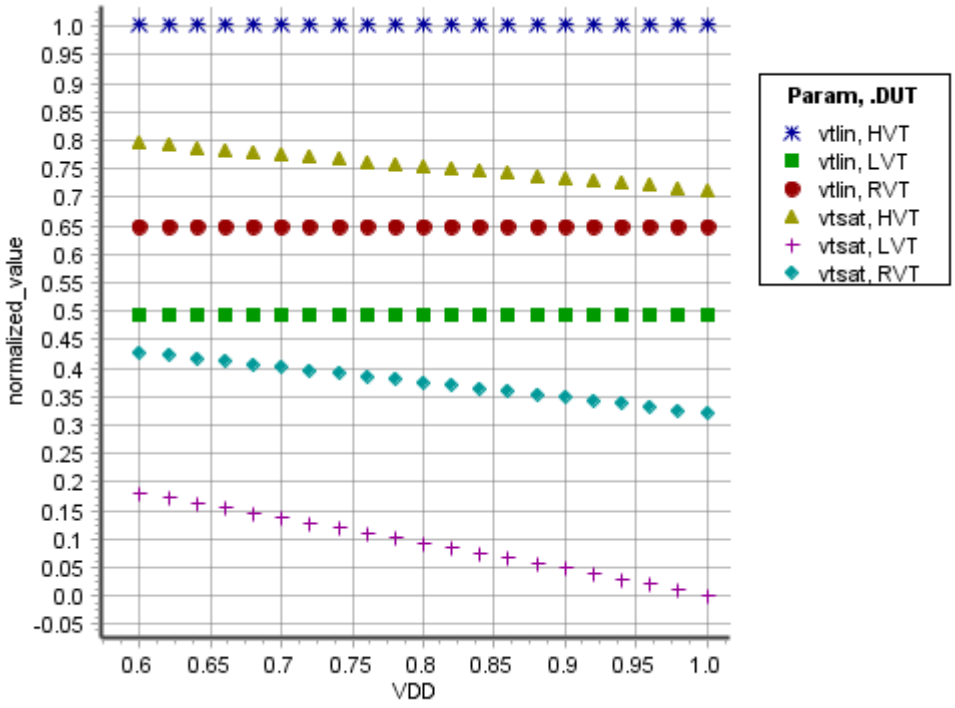

Figure 3.4: NMOS V_T type normalized threshold voltage simulation results.

Once silicon data are available, a full correlation study will be performed to validate the DUTs, DOE, and the technology libraries and models. All errors or departures from expected observations in the simulation and silicon data will need to be studied and root caused to either the layout, circuit design, process, technology libraries, or silicon models. The SLM test structures will eventually be used as a process monitor for the foundry once the early learning is completed. Several secondary effects, such as drain induced barrier lowering (DIBL), will also be studied, as will the spatial and orientation effects of the SLM test structures and DUTs to fully understand the variability in the process across the wafers. This data will eventually drive improvements to the next version of the SLM test structures, which are discussed in the next chapter. These efforts should help drive a robust technology and manufacturing solution for both the foundry and the design house.

Chapter 4: Conclusion and Future Work

The design and implementation of the SLM test structure for 14nm CMOS FinFET technology fulfills the two main goals of the project. The first goal was to create a vehicle for the continuous inline monitoring of FinFET parameter variation at the target foundry at multiple points in the manufacturing line. The unique modular and combinational architectural style of the SLM test structure truly lends itself to the rapid parallel testing of FinFET devices across the wafer. As part of the design, an extensive DOE was also formulated for the FinFET devices across multiple physical attributes of the FinFET device. This solves the problem of tracking the process and manufacturing variation in submicron FinFET technology across multiple device attributes which ultimately affects cost and yield. The second goal was to obtain silicon data for true design-like standard cell implementations of FinFET devices for design modeling as well as yield and performance prediction of the product. This was accomplished by utilizing FinFET standard cell layouts and implementing filler cells to mimic local stress effects and density variation of an actual design as opposed to isolated transistors typically used by foundries. This solves the problem associated with inline test transistors not resembling the design close enough to provide meaningful data.

As part of the design and implementation process, a full simulation run was initiated across multiple corners of the technology library to study all 33 flavors of the FinFET DUTs. This helped to validate the design and create a baseline for the existing technology libraries and models. Additionally, a full inline test plan with measurement algorithms for key parameters like I_{DSAT}, I_{DLIN}, I_{DOFF}, V_{TSAT}, and V_{TLIN} was also created for the test platform at the target foundry. The design and implementation process also highlighted

several avenues for improvement and future work for the SLM test structure. These are addressed in the next section.

4.1 Future Work and Improvements

The immediate next step is to examine the simulation scripts, technology libraries, and modeling files to understand the results of the simulation data with respect to all the physical attributes of the devices. As needed, simulations will be re-run for any significant updates to the library or models. The next step will be for the SLM test structure to be placed on actual wafers for test. The starting point for this step will be for the inline test plan to be presented to the foundry to implement a test strategy to obtain silicon data as soon as possible. Once silicon data are available, the data will then undergo extensive correlation to simulation data. A significant amount of analysis work will need to be devoted to the measurement system as well. Any corrections or modifications to measured data will need to be evaluated and implemented as needed. Once the measurement noise and test precision issues are understood, relevant changes will need to be implemented in the analysis and inline testing. The results of the correlation study and measurement system analysis will be used to update technology libraries and models as needed. As the technology node matures, the inline test plan may change to a process monitoring mode, where test parameter data are collected on a sample basis to track key I-V test parameters as a control mechanism only.

Several improvements and enhancements to the SLM test structure and its usage are already the subject of research. One avenue is exploring the ways in which the test parameter data can be used for an effective wafer pruning strategy for cost reduction [8]. Another avenue is to extract delay or power estimation numbers from the test parameter

data or add additional test parameter measurements [3]. The challenges with measuring leakages by either masking of minimizing other currents either through test or the architecture are also being explored [5]. One example is to incorporate keeper type structures to turn off power or leakage paths to unused sections of the array [10]. New test structures for interconnects or passive devices like resistors and capacitors to the SLM test structure are also being explored. Special test structures designed to measure or rate in-die variation and DFM rules may also be added to the SLM test structure for future technologies.

SLM test structure design and inline testing in general is a ripe topic for innovation and research in the semiconductor industry. It's simple design approach, small size, modular architecture, and ease of testability makes it a key enabler for the development of any technology node. The SLM test structure is an essential vehicle to drive yield and cost reduction for large, custom designs in the deep submicron FinFET process technologies of the future.

Bibliography

[1] Strojwas, A.J., "Cost effective scaling to 22nm and below technology nodes," *Design and Diagnostics of Electronic Circuits & Systems (DDECS), 2011 IEEE 14th International Symposium on* , vol., no., pp.2, 13-15 April 2011.

[2] Liu, T.K., "FinFET History, Fundamentals and Future," *2012 Symposium on VLSI Technology, University of California, Berkeley. Short Course*, 11 June 2012.

[3] Jurczak, M.; Collaert, N.; Veloso, A.; Hoffmann, T.; Biesemans, S., "Review of FinFET technology," *SOI Conference, 2009 IEEE International*, vol., no., pp.1-4, 5-8 Oct. 2009.

[4] Sinha, S.; Yeric, G.; Chandra, V.; Cline, B.; Yu C., "Exploring sub-20nm FinFET design with Predictive Technology Models," *Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE*, vol., no., pp.283-288, 3-7 June 2012.

[5] Borkar, S.; Karnik, T.; Narendra, S.; Tschanz, J.; Keshavarzi, A.; De, V., "Parameter variations and impact on circuits and microarchitecture," *Design Automation Conference, 2003. Proceedings*, vol., no., pp.338-342, 2-6 June 2003.

[6] Ketchen, M.; Bhushan, M.; Pearson, D., "High speed test structures for in-line process monitoring and model calibration [CMOS applications]," *Microelectronic Test Structures, 2005. ICMTS 2005. Proceedings of the 2005 International Conference on*, vol., no., pp.33-38, 4-7 April 2005.

[7] Lefferts, R.; Jakubiec, C., "An integrated test chip for the complete characterization and monitoring of a 0.25µm CMOS technology that fits into five scribe line structures 150µm by 5000µm," *Microelectronic Test Structures, 2003. International Conference on*, vol., no., pp.3-63, 17-20 March 2003.

[8] Yeric, G.; Cline, B.; Sinha, S.; Pietromonaco, D.; Chandra, V.; Aitken, R., "The past present and future of design-technology co-optimization," *Custom Integrated Circuits Conference (CICC), 2013 IEEE* , vol., no., pp.1,8, 22-25 Sept. 2013.

[9] Ketchen, M.B.; Bhushan, M.; Costrini, G., "Addressable Arrays Implemented with One Metal Level for MOSFET and Resistor Variability Characterization," *Microelectronic Test Structures, 2009. ICMTS 2009. IEEE International Conference on*, vol., no., pp.13-18, March 30 2009-April 2 2009.

[10] Shimizu, Y.; Nakamura, M.; Matsuoka, T.; Taniguchi, K., "Test structure for precise statistical characteristics measurement of MOSFETs," *Microelectronic Test Structures, 2002. ICMTS 2002. Proceedings of the 2002 International Conference on*, vol., no., pp.49-54, 8-11 April 2002.

Vita

Mayank Gupta is a Member of Technical Staff at Advanced Micro Devices in Austin, Texas and works as a project lead for the Test Chip Silicon Enablement group focused on silicon test and IP characterization for cutting edge technology nodes. He received his Bachelor's in Electrical Engineering degree in 2005 at the University of Texas at Austin, where he also aims to complete his Master's in Electrical Engineering degree in May 2015.

Email: mike.gupta80@gmail.com

This report was typed by the author.