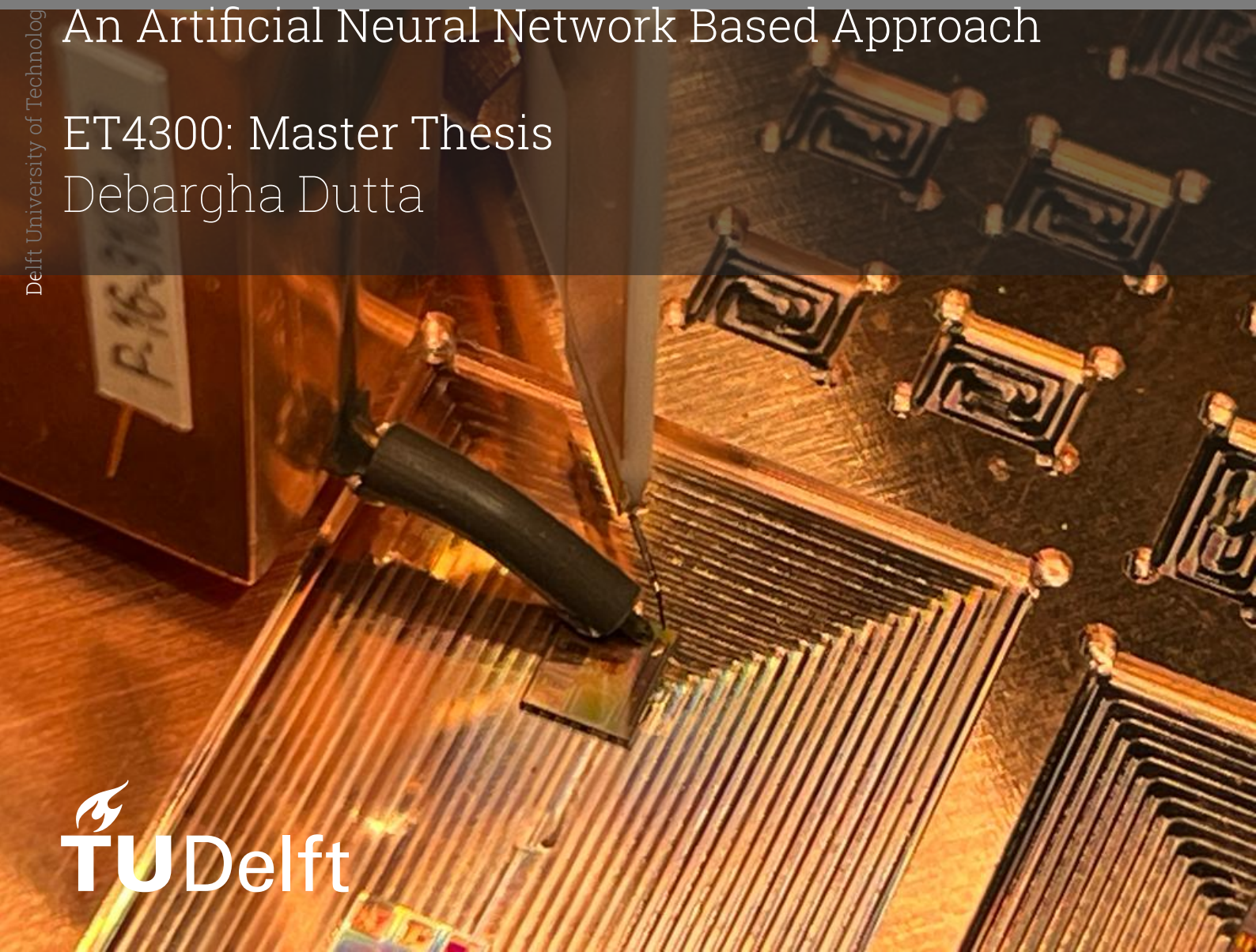
High-frequency Characterization and Modeling of CMOS Transistors at Cryogenic Temperature

An Artificial Neural Network Based Approach

ET4300: Master Thesis

Debargha Dutta

Delft University of Technology



TU Delft

High-frequency Characterization and Modeling of CMOS Transistors at Cryogenic Temperature

An Artificial Neural Network Based Approach

by

Debargha Dutta

Advisor:	Dr. F. Sebastiano
Thesis Committee:	Dr. F. Sebastiano, TU Delft Supervisor
	Dr. M. Spirito
Project Duration:	November 1, 2022 - August 30, 2023
Student number:	5238471

TU Delft

Abstract

To fulfill the vision of scalable quantum computers, cryogenic CMOS is a promising technology to implement the electronic interface for large-scale quantum processors. Research over the years has shown performance improvements in CMOS transistors in commercially viable technology at cryogenic temperatures. While the physics behind the operation of CMOS transistors in cryogenic temperature is known to a great extent, a full-scale compact model is yet not widely available for deep cryogenic temperatures at which qubits operate. For the design of circuits that can be placed in close proximity to qubits, a compatible model needs to be built. In recent years some effort has been made in this direction but most of these are related to DC behavior of the transistor. Very limited work is available that model the RF behavior of transistors at cryogenic temperature across a wide range of bias conditions. As an alternative to a complex physics-based compact model, in this work we have explored the use of artificial neural networks to model the behavior of CMOS transistors in advanced technology nodes. While a non-linear charge-based model using Adjoint-ANN for deep cryogenic temperature has already been shown as an effective approach, we extend the idea of using ANN to facilitate a model of the CMOS transistors. Transistors from 22nm FDSOI technology were characterized for their S-parameter response at room temperature and 4K for this work.

Acknowledgement

First of all, I would like to thank Dr. Fabio Sebastiano, for offering this project to me and for all the supervision during the entire journey.

I would like to extend my regards to Dr. Marco Spirito for being part of the thesis committee. His valuable feedback during the early stages of this work was crucial.

I extend gratitude to Alireza Ghafari and Ehsan Shokrolahzade, for actively helping out during every stage of this work. Special thanks to Ehsan for his help with the calibration, data analysis, and his efforts to identify the problems in the measurement setup.

I am grateful to Niels Fakkel for his regular involvement from the beginning with the probe station. Even with his busy schedule, he was always ready to help.

I would also like to thank my friends here in Delft. It was delightful knowing about different cuisine and language from them.

Finally, special thanks to my family for all their support for the last two years. It would not have been possible without them.

Deb

Delft, August 2023

Contents

1 Introduction

This chapter briefly summarises the status quo in the quantum computing community and the challenges faced in scaling up quantum computers (QC). Further, the methods currently under investigation to tackle such challenges are also described. This chapter will highlight the need to develop transistor models at cryogenic temperature for the design of ultra-low-power circuits. At the end of this chapter, the scope and outline of this work are presented.

1.1. How did quantum computing come about ?

From the establishment of quantum mechanics in the early years of the 20th century to the proposal of quantum computers by Richard Feynman and Yuri Manin in the 1980s till 2023 a lot has been achieved in terms of mathematical modeling of fundamental systems and computing power [2]. Still, it remains a challenge as difficult as climbing Mount Everest barefoot to simulate the behavior of a few interacting particles using quantum mechanics as it will require more computing power than what can be harnessed from conventional computers within a reasonable time frame.

While this naturally draws sarcasm towards the ability of human beings, it also points towards a solution: Build computers whose fundamental operations are based on quantum effects!

The core principle of such a machine is to store information in allowed quantum states and use gate operations to implement algorithms that will solve problems within a reasonable time frame. This idea was firmly established by Peter Shor in 1994 with his algorithm for prime factorization of large numbers.

1.2. Electronics for Large Scale Quantum Computers

Research over the last two decades by the QC community has established the possibility of actually building such systems with small-scale prototypes. But for the practical implementation of quantum algorithms in quantum computers, thousands of qubits are required. Scaling up the number of qubits to such large numbers is an active area of research by various groups across the globe.

It is logical to ask, how will such a system be controlled and how will the information stored be accessed for meaningful human utilization? A solution to this was described in [31]. As shown in figure 1.1, a quantum processor operates at milli-kelvin temperature. Specialized low-power circuits will be placed at 1-4K temperature stage, in close proximity to the processor for control and read-out of qubits. Such circuits can be implemented using commercially viable integrated circuits since these technology platforms allow high-density integration and remain operational even at cryogenic temperatures.

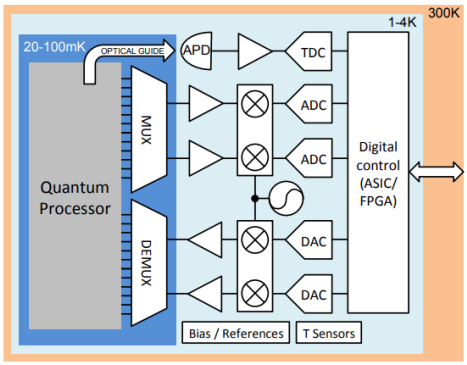


Figure 1.1: Generic electronic platform for the control and read-out of quantum processors. (Image from [31])

Quantum algorithms rely on various quantum gate operations to implement algorithms, just like a classical computer relies on logical gates. To perform such quantum gate operations various types of signals as summarised in figure 1.2 are required. Such signals can be a pulse or modulated high-frequency signals. The intention is to utilize circuit topologies that can generate such controlled signals for biasing or readout of the qubits.

Technology	T_2^*	1-Qubit gate	2-Qubit gate	Qubit read-out	DC-Biasing
Superconducting qubits (Transmons)	2.5 us	~ 6 GHz 20 ns	12ns 40 ns	7-8 GHz, ~ 1 us	flux-bias current
Single-electron spin qubits in a quantum dot	120 us	13-40 GHz ~ 1 us	~ 100 ns		gate voltage
Single-electron spin qubits in a donor system	160 us	30-50 GHz ~ 1 us	~ 100 ns		gate voltage
Singlet-triplet qubit	700 ns	~ 1 ns	~ 1 us		gate voltage
Exchange-only qubit	2.3 us	10 ns ...1 us	Sequence of pulses between different quantum dots		gate voltage
Hybrid qubit	< 10 ns	~ 100 ps	Sequence of pulses between different quantum dots		gate voltage

Figure 1.2: Various signals required to implement various gate operations in different qubit technologies. (Image from [16])

In order to build scalable systems, such circuits will be placed close to the quantum processors so that the requirement for interconnection cables is reduced. The design of CMOS integrated circuits in deep-cryogenic temperatures and often for GHz frequency range introduces challenges. Over the years studies of DC characteristics of CMOS transistors from room temperature to cryogenic temperature have been presented for CMOS transistors in various commercially available technologies. Also, significant effort has been made to investigate and model various physical phenomena like incomplete ionization and interface trapping that are observed at cryogenic temperatures. But till now limited work is available related to the modeling of CMOS transistors for RF integrated circuits. The reasons for considering Cryo-MOSFET models were encompassed in [7]. Integrated circuits for cryogenic applications fabricated on commercially viable foundry processes are usually conservative with carefully drawn design margins. This is mainly due to the lack of process design kits accounting for device behavior at such extreme temperatures. Often multiple design-to-fabrication iterations are required with this approach before a suitable design is available. Power consumption per Qubit is an important factor to be considered while designing a circuit. A quantum mechanical system like a qubit is highly sensitive to temperature. An increase in temperature may destroy the state of the qubits. Since cryogenic refrigerator has limited cooling power, heat generated from the circuit should be under the power budget. In the absence of device models at low temperatures, designers usually have to make conservative design choices as such the circuit topology is usually not optimum.

A compact model that accounts for device physics at such low temperatures can eliminate the need for pessimistic designs. For example [4], dopant freeze-out can be a factor that affects the operation of

a circuit, because of its impact on threshold voltage, and as such local heating can be used to keep the circuit functional. This can be avoided if the cryogenic model of the device is available, enabling the design of alternative circuit topology. Also, the use of super low threshold voltage transistors is worth exploring for such a design. But this will require fabrication units to provide the designers with PDKs that encompass cryogenic temperature effects. Design of integrated circuits is a complicated process that requires multiple groups specialized in various aspects from theoretical research to process engineering. Thus a simplified flow is required that can help foundries to rapidly characterize and build PDKs which can then be evaluated by device physics experts and used by design engineers. This is required to build an efficient feedback loop that builds better insights for process-integration engineers and often optimization of the fabrication process. Extrapolating the room temperature models to cryogenic temperature can lead to over-optimistic CMOS performance. This is definitely not feasible especially due to the ultimate goal of integrating control electronics and qubits on the same substrate.

1.3. Cryogenic MOSFET Modelling

BSIM, EKV, PSP and L-UTSOI are the four major MOSFET compact models that have been used through the years for most commercial processes. These compact models have been developed over the years to accurately model device behavior and are mostly used in the industrial temperature range. But as the demand for integrated circuits for quantum computers became popular, the requirement for cryogenic transistor models also got prominence. As such efforts have been made to use these standard compact models to capture device behavior in cryogenic temperature.

In [34] and [37], a BSIM-based CMOS RF compact model was proposed for 40nm low-power CMOS technology. In [34], the temperature dependence of S-parameter measurement was used to extract the model parameters. The simplified design-oriented EKV model was used to build a cryo-model for 28nm FDSOI technology in [6]. In [25], MOS11/PSP model was modified to build a cryo-compatible compact device model for bulk CMOS technology at 40nm and 160nm nodes. A physics-based cryogenic model compatible for FDSOI technology has also been proposed in [3] based on L-UTSOI from CEA Leti. Compact modeling of n-channel bulk FinFETs using BSIM CMG was proposed in [22]. These physics-based equations can be very rigorous to generate models.

[12] and [23] used a different approach where a small-signal equivalent circuit was used to model the RF behavior of 22nm FDSOI transistors. This approach is different from the previous ones as physics-based equations were not used in these two works. In [24] though, a different approach was presented. Using ANN, the DC IV characteristic was modeled while Ad-joint-ANN was used to model terminal charge functions from terminal capacitance which was captured from the S-parameters of the device. The quasi-static approximation has been made in this approach.

1.4. Thesis Objectives

In [24], an ad-joint ANN-based modeling tool from ICCAP was used to model a CMOS transistor. Using S-parameter simulations, the capacitance between the device terminals was included in the model. Since measured data at cryogenic temperature was not available hence, room temperature simulation data was used for s-parameters. The use of Artificial Neural Networks allows the automatic generation of cryo-CMOS simulation models without requiring any physics-based device modeling effort. This work further extends this idea.

The objective of this work is as follows. First, CMOS transistor S-parameter measurements at room temperature and 4K temperature at various bias points are performed. From the measured S-parameters of the transistor, extract the parameter values of the transistor small-signal equivalent circuit. Using the ICCAP ANN tool generate a model that will map the parameter values to corresponding bias voltages, transistor size, and ambient temperature. Finally, implement the small-signal equivalent circuit using the ANN-generated model in Verilog-AMS language.

1.5. Thesis Outline

In this direction, we start with a review of previous DC and RF characterization literature. Such background data is necessary to understand device behavior and ratify the device measurements. This is presented in Chapter 2.

In Chapter 3, a suitable MOSFET small signal model that can replicate the S-parameter response of

the device in 100 MHz to 40GHz is studied. The usability of the ICCAP ANN tool is introduced.
In Chapter 4, the chip used for S-parameter characterization is described. Also, SOLR calibration and two-step de-embedding is reviewed.
In Chapter 5, the measurement setup is described and measurement results are presented. Practical challenges with data acquisition especially at cryogenic temperature are discussed. The usability of the ANN tool in the model generation process is emphasized here.
Chapter 6 summarizes the findings of this work and presents some future topics for investigation.

2

Review of MOSFET Characterisation at Cryogenic Temperatures

This chapter revisits the cryogenic behavior of MOSFET from TSMC 40nm and Global Foundries 22nm technology. Significant efforts have been made to characterize the DC and RF behavior of these technologies. Since the MOSFET behavior has bias dependency, for any high-frequency analysis it is important to characterize its DC behavior. The terminal capacitance that forms the charge-storing element of the device varies with DC bias voltage in addition to trans-capacitance and channel resistance. These parameters in turn become the deciding factor for the RF performance like maximum oscillation frequency and unity gain frequency.

2.1. Review of DC Characteristics

The most important properties of interest from DC characterization are the sub-threshold swing, threshold voltage, mobility, and kink effect. These parameters are used to compare the change in behavior of the transistor as the temperature goes from room temperature to 4K. The kink effect is not observed in these technology nodes due to the high vertical field. Sub-threshold swing is defined as V_{gs} increase required to raise the drain current by a decade for devices biased with V_{gs} below the threshold voltage. Inverse of it, sub-threshold slope increases as the temperature goes low. A steeper sub-threshold slope is observed at cryogenic temperatures means implies more ideal switching behaviour of the devices. At lower temperature mobility improves due to reduced phonon scattering till about 40K after which it tend to saturate due to Coulomb scattering which starts to dominate. The increase in mobility means more drain current for the same over drive voltage compared to room temperature as such resulting in faster switching.

Figures 2.1, 2.2 and 2.3 show the current vs terminal voltage variation at temperature for an NMOS transistor from TSMC 40nm technology. While a higher drain current can be observed with gate voltage, an increase in threshold voltage is observed. From a device measurement perspective, figures 2.1 and 2.2 can be used to qualitatively analyze the correctness of device data especially for the experimental setup for this work. The reason for this is discussed in Chapter 5. The threshold voltage calculated using the extrapolation-in-linear-region method is shown in figure 2.4. It increases linearly as the temperature decreases to about 50K and thereafter saturates.

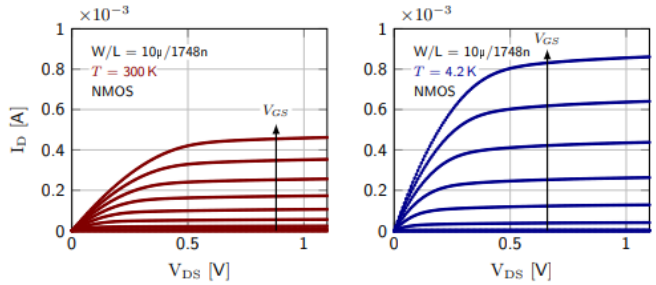


Figure 2.1: $I_D - V_D$ curves for TSMC 40nm NMOS[1]

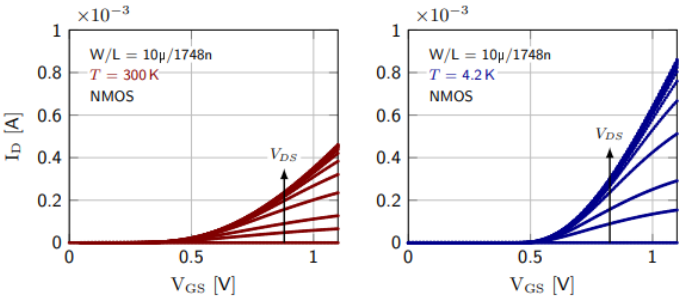


Figure 2.2: $I_D - V_G$ curves for TSMC 40nm NMOS[1]

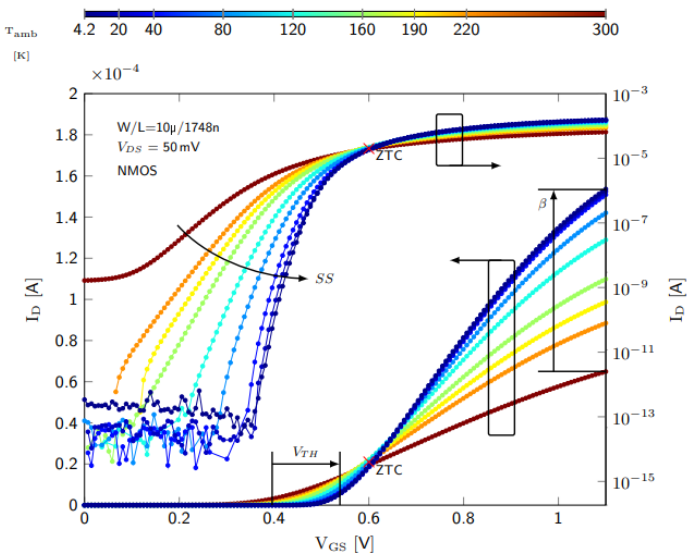


Figure 2.3: $I_D - V_G$ curves biased at V_{DS} = 50mV as function of temperature for TSMC 40nm NMOS [1]

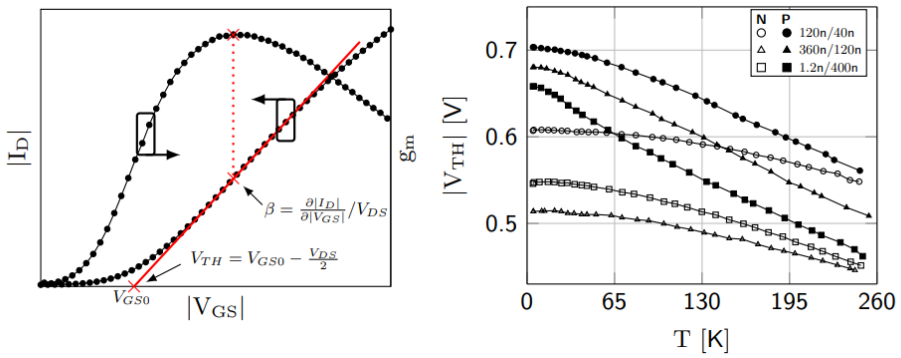


Figure 2.4: Threshold voltage behaviour. Left: ERL method for threshold voltage extraction. Right: Extracted V_{TH} as a function of temperature for NMOS(N) and PMOS(P) (image from [1])

Similar work was presented in [12] for 22nm FDSOI technology node from Global Foundries. These are shown in figure 2.5. From the plots, it can be observed that threshold voltage of both n and p-type MOSFETs show a linear increase with decreasing temperature and tends to saturate at around 70K. This behavior is similar to 40nm technology transistors. In addition, the $I_D - V_{GS}$ curves at $V_{DS} = 50mV$ for varying temperatures suggest that the ZTC is around 600mV for both cases. The ZTC is explained as a simultaneous effect of temperature-dependent mobility and threshold voltage. A detailed explanation of ZTC for FDSOI technology can be found in [10]

As temperature sweep is beyond the scope of this work and from a more practical measurement perspective the IV curves at room temperature and lowest cryogenic temperature reported are of interest.

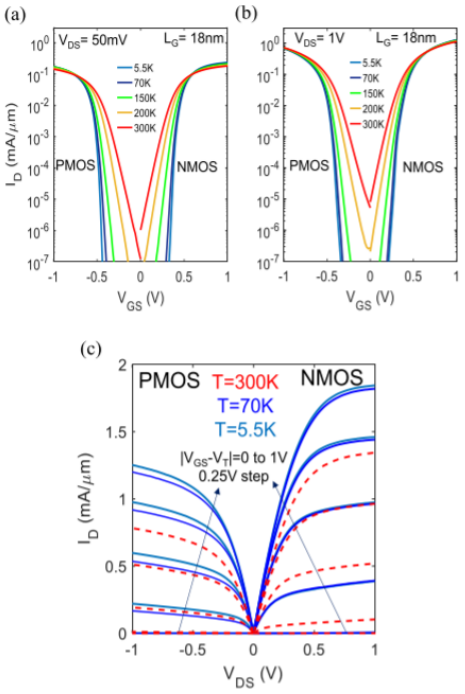


Figure 2.5: IV characteristics of Global Foundries 22nm FDSOI NMOS(N) and PMOS(P) (image from [12])

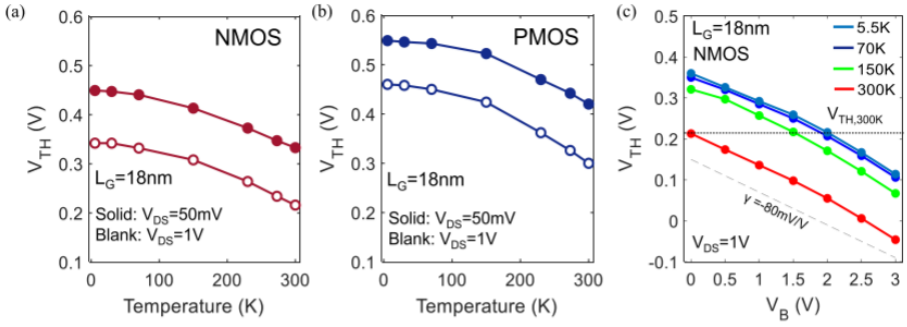


Figure 2.6: Threshold voltage variation with temperature and bias-biasing voltage (image from [12])

2.2. Review of RF Characterisation

While DC characterisation gives the current versus voltage relation for a device, it is not sufficient to describe the entire operation of transistors. This is because of the presence of capacitive behaviour, effect of which starts dominating as the frequency of operation increases. A representative diagram of the cross-section[11] is shown in figure 2.7.

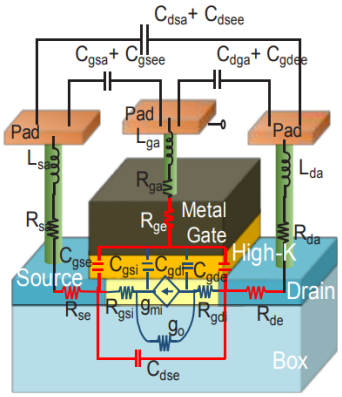


Figure 2.7: Cross section view of FDSOI MOSFET with small signal equivalent circuit schematic (image from [11])

Based on the above figure FDSOI transistor can be split into three parts. The first part is the external access paths which is defined by the parasitic resistances R_{sa}, R_{da}, R_{ga}, and the parasitic capacitance between metal pads. The second part is formed by extrinsic parasitic resistance R_{se}, R_{de} and R_{ge} and capacitance C_{gse}, C_{gde} and C_{dse}. The last part is the intrinsic FET parameter set. This includes the trans-conductance (g_{mi}), trans-conductance (g_{mi}), output conductance (g_o), and intrinsic terminal capacitance (C_{gsi},C_{gdi},C_{dsi}).

Interestingly, it is possible to model the MOSFET in FDSOI in alternative ways. In [23] (see figure 2.8), a simplified small-signal model was used to model the transistor for the same technology node. In this approach, the buried oxide layer is considered by including three capacitors from the gate, drain and source terminals to the bulk. The well underneath the oxide layer is represented with resistance. Note that unlike [11], the drain-to-source capacitance is not there in this model, instead drain-bulk capacitance is used. While both [11] and [23] are effective in capturing the device behavior, the former can be viewed as a source reference model. On the contrary, the later model is a bulk-referenced model. It is interesting to note the bulk-referenced model was originally proposed for bulk CMOS transistors.

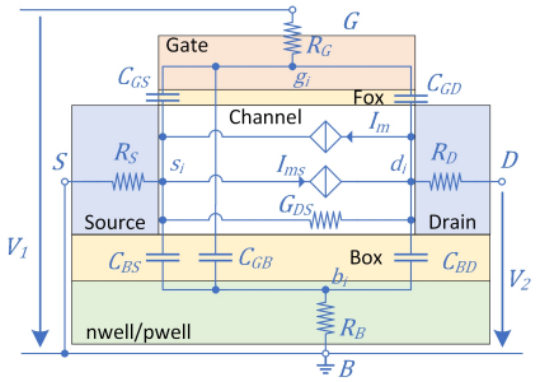


Figure 2.8: Simplified small signal model of MOFET in 22nm FDSOI technology (image from [23])

From the perspective of modelling at cryogenic temperature, it is important to know the variation of the various model parameters with temperature. For the model proposed in [11] and [12], it was found that extrinsic capacitance does not vary significantly while extrinsic resistance decreases with temperature due to a decrease in sheet resistivity. Although the reason for constant extrinsic capacitance was not mentioned explicitly by the authors. Looking at the figure 2.7 and use of cold FET de-embedding method to remove the extrinsic parasitic network, these are used to model overlap capacitance between gate and source, gate and drain, and the drain-source capacitance through the buried oxide layer in the model and are as such temperature insensitivity.

Similar behavior can be observed for the model in [23] and is shown in figure 2.9. The values of gate-bulk capacitance C_{GB} and trans-capacitance C_{ms} are only calculated at 300K. The authors have considered the bulk resistance (nwell/pwell) as zero for lower temperatures due to reduced phonon scattering and no dopant freeze-out effect on bulk. C_{GB} and C_{ms} are coupled to R_B in the Y-parameter equations. If R_B is considered zero, they cannot be calculated from the equations presented.

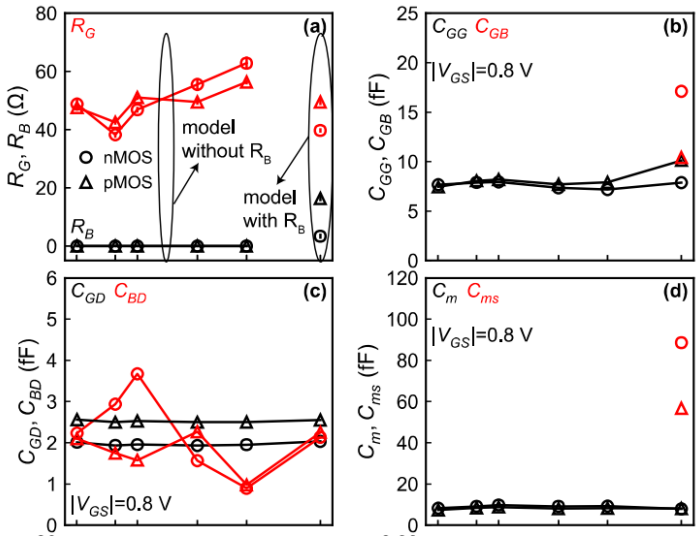


Figure 2.9: Capacitor and resistor variation with Temperature at $V_{GS} = 800mV$(image from [23])

For RF model of MOS transistors unity-gain frequency (f_T) and maximum-oscillation frequency (f_{max}) are commonly used performance metric. As such they can also be used to validate the model.

For determination of f_T the short circuit current gain H_{21} is calculated from device S-parameters. Then a linear extrapolation is made to reach unity gain frequency. Usually, the transistor is biased at peak trans-conductance gate bias with drain bias in saturation.

Similarly, linear extrapolation of unity power gain is used to determine f_{max}. The bias condition is kept the same as in the case of f_T. Results of such analysis were presented in [23] (figure 2.10),

where the peak values of f_T and f_{max} were determined at different temperature. Although no standard globally accepted method, that can used to compare models, is available as of date; RF performance metrics is often used to test the accuracy of the extracted model. Although this definitely is not a conclusive way to rank any two models.

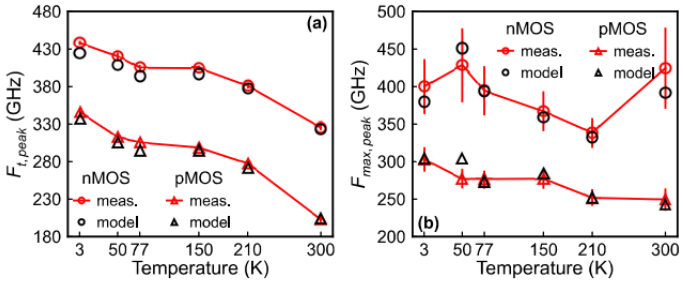


Figure 2.10: f_t and f_{max} variation with Temperature (image from [23])

The transit frequency and maximum oscillation frequency for the equivalent circuit in 2.8 were derived in [13].

$$f_t = \frac{g_m}{2 \times \pi \sqrt{C_{gg}^2 - (C_{gd} + C_m)^2}} \tag{2.1}$$

As temperature decreases, C_{gg}, C_{gd} and C_m stay relatively constant. But due to an increase in mobility, g_m increases. The expression for f_t thus explains the increasing trend observed at lower temperatures.

The expression for f_{max} is more complicated. It is calculated using Mason's unilateral gain expression by equating it to 1.

$$f_{max} = \frac{\sqrt{\omega_{p2}\left(\sqrt{4 \cdot K + \omega_{p2}^2} - \omega_{p2}\right)}}{2\sqrt{2}\pi} \tag{2.2}$$

where

$$K = \frac{g_m^2}{4 \cdot (C_{gb}R_b\,(C_{gb}G_{ds} - C_{db}g_m) + C_{gg}R_g\,(C_{gd}g_m + C_{gg}g_{ds}))}$$

$$\omega_{p2} = \sqrt{\frac{C_{gb}R_b\,(C_{gb}g_{ds} - C_{db}g_m) + C_{gg}R_g\,(C_{gd}g_m + C_{gg}g_{ds})}{R_bR_g \cdot (C_{db}C_{gg} + C_{gb}C_{gd}) \cdot (C_{gg}\,(C_{db} - C_m + C_{ms}) + C_{gb}\,(C_{gd} + C_m))}}$$

From the expression of f_{max} it is somewhat non-intuitive to explain the trend. Although, due to an increase in g_m while other parameters stay approximately constant with temperature, one can expect an increasing trend with decreasing frequency. But in figure 2.10, it is observed that f_{max} increases from 210K to 50K and then decreases.

In [34], RF characterization was carried out for 40nm CMOS technology. In this case, BSIM compact model was extended to 6K. f_t showed a similar increasing trend as for FDSOI with a decrease in temperature. This is shown in figure 2.11. f_{max} was not reported in this paper.

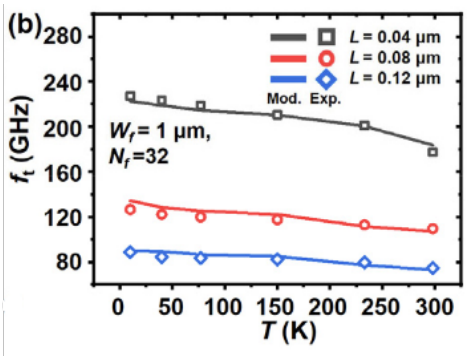
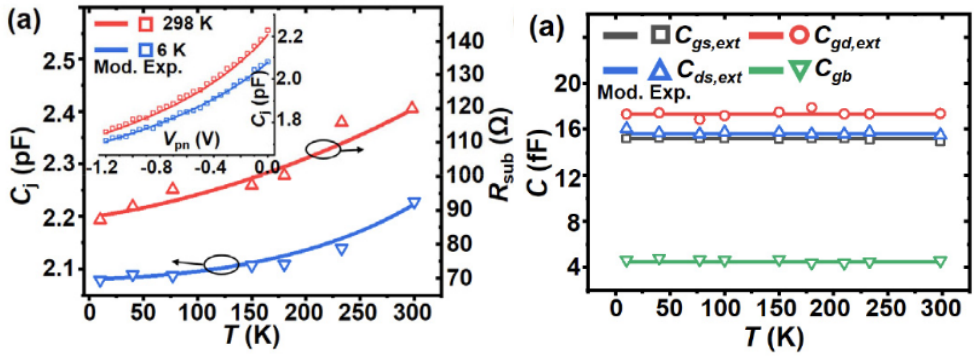


Figure 2.11: f_t variation with temperature (image from [34])

Figure 2.12 shows the variation of the junction capacitance, substrate resistance, and extrinsic capacitance variation with temperature. The diffusion-type junction capacitance at the source and drain junctions with the bulk shows a monotonic decrease with the temperature drop. The reason for this decrease was justified by an increase in depletion width due to the incomplete ionization of the bulk well. The substrate resistance decrease at low temperature. R_{sub} includes the diffusive resistance of the well and surrounding metal connections. The decrease was attributed to low resistivity metal connections and degenerate doping of the source and drain region



(a) Junction capacitance and Substrate resistance variation with temperature **(b)** external capacitance variation with temperature

Figure 2.12: Junction capacitance and Extrinsic capacitance variation with temperature for HLMC 40nm process [34]

The increase in f_t can be justified in a similar way as before. The g_m increases, external capacitance remain the same, and junction capacitance is reduced. This will boost the transit frequency at low temperatures.

2.3. Summary

Recent work related to DC and RF characterization of CMOS transistors in bulk CMOS and 22nm FDSOI technology nodes was reviewed. As the DC bias point is an integral part of analog RF operation of transistor temperature-dependent DC behavior was reviewed. Changes in drain current, trans-conductance, and shift in threshold voltage with temperature were seen. RF characterization using S-parameters reveals the terminal capacitance of the device. Variation of such capacitance with temperature was reviewed based on recent work. Maximum oscillation frequency and transit frequency are commonly used RF figures of merit. As temperature decreases increase in g_m causes f_t and f_{max} to increase.

3

Study of the CMOS transistor small-signal model

So far we presented the need for a compact model at cryogenic temperature for CMOS transistors for scaling up Quantum Computers. We briefly reviewed various efforts made in this direction. Most of the work presented till now focuses on the DC characterization of transistors in cryogenic temperature. In recent years, an extension of compact models from BISM, PSP and EKV to cryogenic temperature has been published. Limited effort has been made to develop cryogenic compact models that reliably include RF behavior. Recently, RF characterization at cryogenic temperature has been published. A simple small signal equivalent circuit was used to model the transistor in cryogenic temperature. Use of ANN-based non-linear model development assuming quasi-static operation of the transistor has also been reported. But overall very limited work can be found in terms of RF characterization of transistors in cryogenic temperature. Note that due to the large number of parameters in standard compact models extending them to cryogenic temperature or building physics-based models is a time-consuming rigorous process. Possibly this can be the reason why limited progress has been reported with these traditional approaches. On the other hand, a small signal model of transistors can be used to capture the linear quasi-static operation of CMOS transistors for a fairly wide range of frequencies which is required for AC analysis. It can be directly extracted from S-parameter measurement and as such is very useful to model transistor behavior even at cryogenic temperature.

3.1. Function generation Using ICCAP ANN Tool

Although such physical models are extremely important especially when process optimization is under consideration, from a circuit design perspective a model is required that can determine the electrical behavior of the device as accurately as possible. Instead of having the physical model equations which in most cases are cumbersome to develop, the use of an artificial neural network to generate functions that map the terminal voltages, width, length, and device temperature to the variation of the parameters can be used in the device model back-end in the circuit simulator. In addition, Verilog-AMS language can be used to define the behavior of the device operation using these parameters. This model development flow is explored in this project.

The usability of ANN for model generation was demonstrated in [24] and [26]. The basic requirement for this approach is the model parameter and the independent variables should have a dependency. If this holds true, a meaningful function can be generated by ANN training.

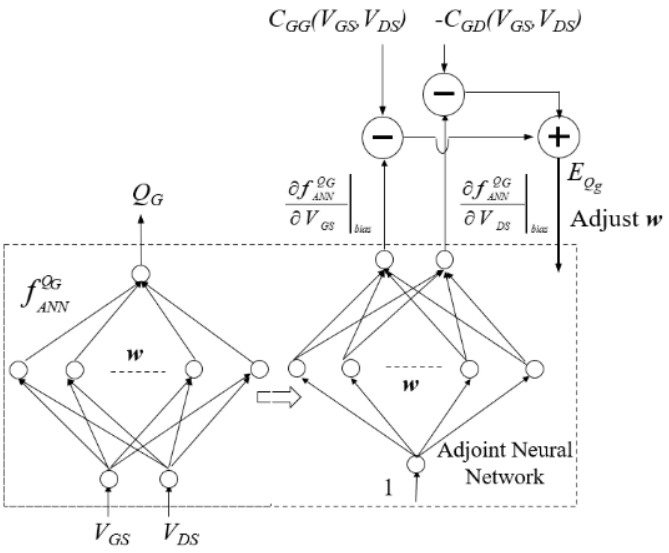


Figure 3.1: Model training for gate terminal charge function from C_{GG} and $-C_{GD}$ data from S-parameter of device [24]

Figure 3.1, a representative diagram is shown for model training of gate terminal charge. The capacitance seen in the gate terminal is C_{gg} when excitation is from the gate terminal. While $-C_{gd}$ is the capacitance seen from the gate terminal when excitation is from the drain terminal. Since these capacitance values are non-linear such a non-linear charge model will be required to model these. The Adjoint ANN will start from an initial point for charge function Q_g. The partial derivative of this charge function is compared with the capacitance provided as training input. The error is used to adjust the weights and bias in the ANN. This is a top-level description of the model.

The DC bias currents at the gate and drain terminals can also be modeled using the regular ANN training method in the tool; adjoint-ANN training is not required in that case. If a quasi-static operation is assumed for the transistor, using the DC terminal current model and the non-linear charge model, a MOSFET can be implemented with reasonable accuracy.

The accuracy of this method is dependent on the data set used for training. Depending on the complexity of the variation of the parameter with the independent variables, the number of hidden layers and neurons per layer is to be selected such that the function fits the data points. Note that in reality the data from measurement is expected to be scattered about a best-fit line. If a very large size of ANN is selected it will reduce the training error by fitting the equation to all the scattered points. The selection of ANN size is an iterative process and sometimes takes a long time before the optimum is reached. Also, the model presented in [24] does not give accurate results for small signal analysis. One can simulate the Y-parameter with the model where the imaginary part will be replicated but for the real part of the Y-parameter, the frequency dependency will be absent. Thus ANN modeling approach has its own challenges too.

3.2. Review of Small Signal Model for MOSFET

In a previous discussion, it was mentioned that the small-signal model of transistors is useful for the AC analysis of circuits. While a compact model is an ultimate requirement, it is not a straightforward thing to do especially at cryogenic temperature due to device physics complexity. The ANN modeling approach already gives us a method to model transistors in quasi-static operation that can be used for circuit simulation with reasonable accuracy but has limited capability in terms of SP or AC simulations as it does not account for the variation of the real part of Y-parameters with frequency. This will over-estimate the RF figure of merits like f_t and f_{max}. In this work, the ANN modeling approach is used to model a small signal equivalent circuit of the transistor with a target frequency of up to 40GHz. For this purpose, we hereafter briefly review the various small signal model available in publications. A simplified model is preferred that can be easily implemented for small signal analysis across bias range. In the previous chapter, two recently used proposed small signal equivalent model for FDSOI technology was presented. RF characterization results from measurements were mapped to these

equivalent circuit parameters from 300K at cryogenic temperature. These two models are shown here for ready reference.

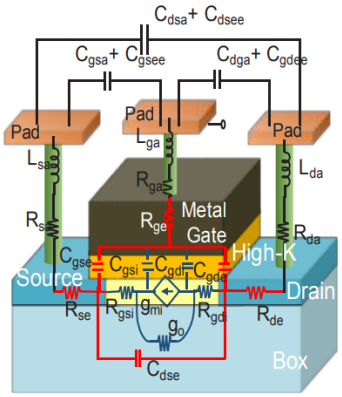


Figure 3.2: Cross section view of FDSOI MOSFET with small signal equivalent circuit schematic (image from [11])

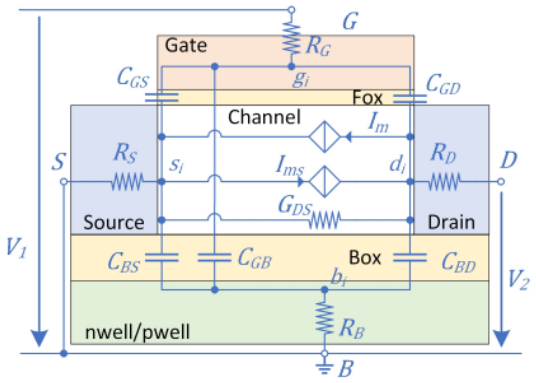


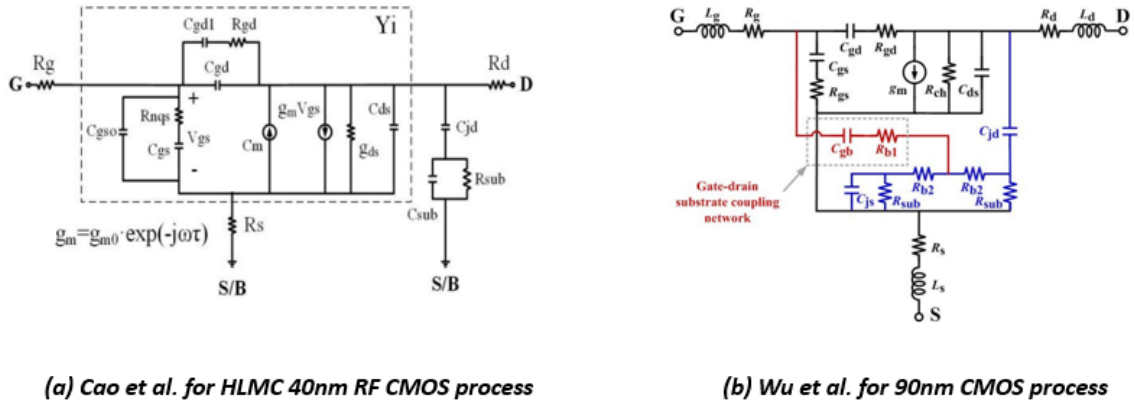
Figure 3.3: Simplified small signal model of MOSFET in 22nm FDSOI technology (image from [23])

To extract the parameters from the circuit shown in figure 3.2, first, the S-parameters of the device under cold FET conditions are used to de-embed the extrinsic network. The intrinsic FET parameters are then derived from these de-embedded S-parameters. In this model, the effect of the buried oxide layer and the well underneath has been ignored. But a good agreement was till 35 GHz was reported in this work.

In figure 3.3, a simplified bulk-referred model which was originally proposed for the bulk-CMOS process was used to model the same technology node til 3.3K. The key highlight of this approach is the inclusion of the buried oxide layer effect using the terminal to intrinsic bulk capacitance. The simplifications suggested with this model firstly that the source and drain access parasitic left after de-embedding the pad parasitic can be ignored as they are very small numbers. This was also stated in [11]. The effect of C_{sb} was also ignored as it does not appear in the Y-parameter expressions (stated later in the next section) till the second-order frequency terms. One difference between this and the previous model is that in this case, the trans-conductance model includes the phase factor associated with it due to non-quasi-static response at very high frequencies. In [13], this has been elaborated. Also, the corresponding Y-parameter expressions are such that they can be solved analytically in a few simple steps and this has been presented in[13]. Thus, the need for doing such an analysis is not required. Also, note that this model is to some extent flexible as it was originally used to model bulk CMOS transistors.

The presence of the substrate in bulk-CMOS technology like the TSMC 40nm or HLMC 40nm requires the equivalent circuit to account for the effect of the substrate at higher frequencies. Although the fundamental principle remains the same as discussed earlier, analysis of the Y-parameters in terms

of the model parameters gets tedious. Figure 3.4 shows the small signal equivalent circuit schematic used to model the RF behavior transistors of bulk-CMOS technology in [9] and [38].



(a) Cao et al. for HLMC 40nm RF CMOS process

(b) Wu et al. for 90nm CMOS process

Figure 3.4: Bulk CMOS transistor Small Signal Equivalent Circuits proposed for RF characterization and modeling

Both of these models require significant effort to extract the parameter values at each bias point. Although the intrinsic transistor and substrate are modeled in a different manner, one similarity is the modular nature. Regarding admittance parameters, it is possible to visualize the intrinsic FET and the substrate network as separate units. This is helpful in the derivation of the functions relating Y-parameters to the circuit parameters. Also, these models exploit the zero-bias condition of FET to extract extrinsic bias-independent parameters as a first step in the parameter extraction process.
It should be noted that methods proposed in [9] and [38] rely on frequency approximations in the measured frequency range of 0 to 40GHz to estimate various parameters. Such approximation while may seem suitable for an analytical approach to calculate the parameters; can in principle add complexity in reaching the optimum values. [38] approximates 10GHz as a low frequency for model simplification for parameter extraction. While this might have worked, such assumptions are not valid in general; as such repeatability and usability with other bulk CMOS processes may be difficult if not impossible.

Another approach to model 40nm bulk CMOS was presented in [13]. This approach uses the same simplified equivalent circuit as in figure 3.3. Since the operating frequency range of interest is much less than transit frequency, the assumptions made for simplification in this approach are more intuitive and repeatable with different technologies. Of course, some deviation can be observed but these should not be drastic.

In previous paragraphs, four different small signal equivalent circuits were reviewed. Out of these, the one presented in [23] is used in this work due to its simplicity and ability to include the phase associated with trans-conductance and proven use in cryogenic temperature. While [23] used the model in strong inversion and saturation, we analyze it in different regions of operation. Further, we use the capability of ANN to develop a model out of the extracted parameters as described before which will enable small-signal model simulations across the entire bias range and up to 40GHz frequency in cryogenic temperature.

3.3. Description of the used small signal equivalent circuit

As the frequency of operation goes higher (the highest frequency will be 40GHz in this work), the extrinsic part of the transistor,i.e., the region outside the channel region, will start contributing to the operation of the device. The model should be able to account for such non-ideality. It is possible to build a small signal quasi-static equivalent circuit of an RF MOSFET with great detail. But this can make the analysis complex and extraction of individual elements of the model shall require complex optimization algorithms. This implies there should be a trade-off between the complexity and accuracy of the model. Also, the layout of the transistor can introduce additional parasitic elements. Considering these factors a simplified circuit with a reduced number of parameters is a rational choice to begin with.

Depending on the operating frequency range, deviations from measurement can be accounted for by considering elements that were previously not considered.

We investigate the quasi-static small-signal equivalent circuit for RF MOSFET shown in figure 3.5. As seen earlier, this model can be used both for FDSOI and bulk-CMOS. The capacitors C_{GS}, C_{GD}, and C_{GB} shown in the figure include the overlap capacitance, fringing capacitance, and intrinsic capacitance. C_{SB} and C_{DB} are the summation of intrinsic and junction capacitance. I_m and I_{ms} are two voltage-controlled current sources that model the ac current due to the trans-conductance effect from the gate-source and source-bulk potential and are defined as:

$$I_m = Y_m \left(V_{gi} - V_{bi}\right) \tag{3.1}$$

and

$$I_{ms} = Y_{ms} \left(V_{si} - V_{bi}\right) \tag{3.2}$$

Note that the model is using the bulk as the ground reference. The controlled current sources have a complex gain formed by trans-conductance and trans-capacitance. Such that $Y_m = G_m - j\omega C_m$ and $Y_{ms} = G_{ms} - j\omega C_{ms}$. The definition of G_m, G_{ms}, C_m and C_{ms} is based on the first-order approximation of the non-quasi-static operation of the intrinsic transistor [17]. CMOS transistor when operating at frequencies $\ll f_T$, the frequency dependence of trans-admittance can be accounted by:

$$Y_m = g_m \left(1 - j\omega\tau_{qs}\right) \tag{3.3}$$

$$Y_{ms} = g_{ms} \left(1 - j\omega\tau_{qs}\right) \tag{3.4}$$

where τ_{qs} is the transit time of the device and is bias dependent. A detailed explanation of this can be found in [35],[29]. A direct consequence of the above two equations is that the time-varying signal across gate-bulk and source-bulk terminals (since the reference is bulk in this approach) will have a dominant trans-conductance effect at low-frequency operation. As the frequency increases, the trans-capacitance effect will contribute in parallel. As a result of this, the capacitive effect of the gate on the drain terminal will be $C_{GD} + C_m$. A similar argument holds for source and bulk terminals using C_{ms}.

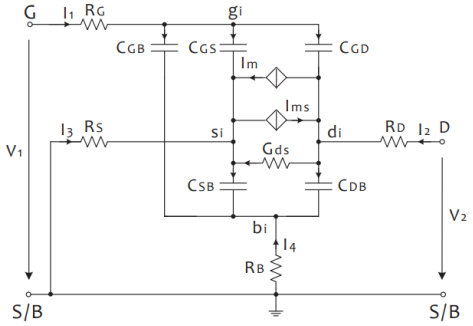


Figure 3.5: Quasi-static Small Signal Equivalent Circuit proposed for RF MOSFET (schematic from [13])

The Y-parameter analysis of the schematic shown in figure 3.5 can be convoluted and not lend itself to easy extraction of the different circuit parameters. Hence some simplifications are necessary. The source and drain terminal resistances can be neglected as the poles due to these two being typically above transit frequency [14],[18]. C_{SB} also does not show up in the first-order analytical model [13] as the source is tied to ground and source-bulk voltage drop is negligible. This simplifies the schematic as shown in figure 3.6 below.

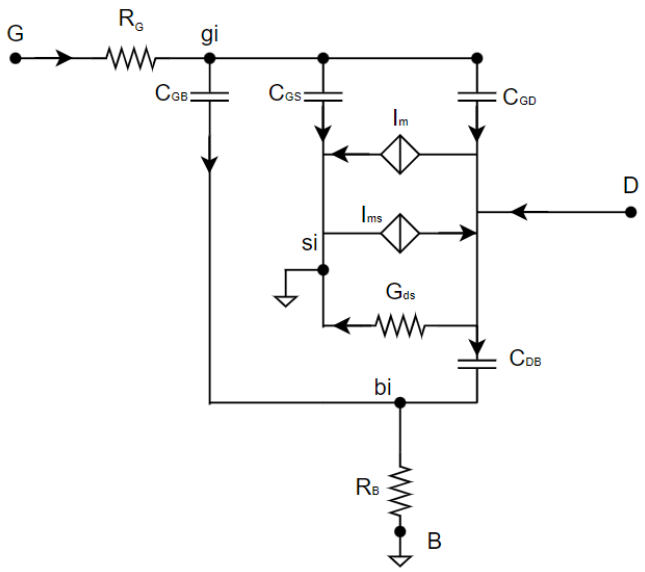


Figure 3.6: Modified Small Signal equivalent circuit

The intrinsic FET and the substrate can be seen as two subsections in parallel. The substrate network is formed by C_{DB}, R_B and C_{GB}. So their Y-parameter can be analyzed separately and then summed up. Further, the gate resistance is in series to this combined network; as such their Z-parameter matrix can be added. Finally, neglecting the higher-order terms, as was shown in [13], in the Y-parameter matrix, a simple set of analytical expressions describing the small signal equivalent circuit was obtained as shown below.

$$Y_{11} \approx \omega^2(C_{GB}^2 R_B + C_{GG}^2 R_G) + j\omega C_{GG} \tag{3.5}$$

$$Y_{12} \approx \omega^2(C_{BD}C_{GB}R_B - C_{GD}C_{GG}R_G) - j\omega C_{GD} \tag{3.6}$$

$$Y_{21} \approx G_m + \omega^2(C_{GB}R_B(C_{BD} - C_m + C_{ms}) - C_{GG}R_G(C_{GD} + C_m)) - j\omega(C_{GD} + C_m) \tag{3.7}$$

$$Y_{22} \approx G_{DS} + \omega^2(C_{BD}R_B(C_{BD} - C_m + C_{ms}) - C_{GD}R_G(C_{GD} + C_m)) + j\omega(C_{BD} + C_{GD}) \tag{3.8}$$

The ten unknown terms in these equations can be determined analytically without the requirement of any non-linear regression algorithm or any further assumptions in frequency, as shown in the following section.

3.4. Simulation based analysis

The proposed model is tested against simulation data from Global Foundries 22nm FDSOI technology PDK at room temperature. Since this technology PDK is commercially used for circuit design in the industrial temperature range, it is a good reference for such a study. S-parameters from 100 MHz to 40GHz at various DC bias conditions are simulated for transistors from this technology. These are converted into Y-parameters and the analytical extraction method as in [13] is used. The model is compared with the PDK model for a large frequency range. This comparative analysis indicates the operating region limitation of the proposed small signal equivalent circuit.

3.4.1. Parameter Extraction

From the general expression of Y-parameters presented in the previous section, the following steps are followed sequentially to determine the parameters.

1. Determine gate-drain capacitance C_{GD} from imaginary part of Y_{12}.
2. With C_{GD} known, from imaginary part of Y_{21} and Y_{22} determine C_m and C_{BD} respectively.
3. Extrapolate the real parts of Y_{21} and Y_{22} to zero to obtain G_m and G_{DS}. These values can also be determined from DC I_{DS} vs V_{GS} and I_{DS} vs V_{DS} curves, obtained from DC sweep measurements. Although, it is suggested to extract these from the Y-parameter so that the parameter values stay consistent with self-heating considerations, especially for cryogenic temperatures where channel heating can have a significant impact due to the longer time required for S-parameter measurement over the wide frequency range of measurement from 100 MHz to 40 GHz. Only if the DC data is acquired in parallel with S-parameter measurement, it can be used to determine these G_m and G_{DS} as both the measurements will be in similar channel temperature conditions.
4. From Y_{11}, determine C_{GG} = $C_{GS} + C_{GD} + C_{GB}$.
5. Note that the frequency-dependent real parts of the Y-parameters form a set of four equations. The four unknown variables C_{GB}, C_{ms}, R_G and R_B can be determined. In order to do so, use the derivative with respect to the frequency of the real part of y-parameters to generate four equations. Note that due to the quadratic nature of these equations, two sets of solutions will be generated. A condition statement is thus required to select the real positive values. The calculated parameters can be negative in value for certain bias points. Also, note that the measured Y-parameters is expected to be noisy. Thus, the best-fitting curve is determined in such cases. It can be expected that at certain bias points the parameters might take values that are not physically meaningful such as negative values or very large values.
6. The parameter values for each bias point should be able to fit all the Y-parameters with minimum error for the entire frequency range. Although no standard methodology to define the quality of fitting is available, one method can be to look at the root-mean-square error over the entire frequency range. Thus, we can write

$$e_{ij} = \frac{Y_{ij,measured} - Y_{ij,model}}{Y_{ij,measured}}$$

$$e_{ij,rms} = \sqrt{\left(\frac{1}{N}\sum eij^2\right)}$$

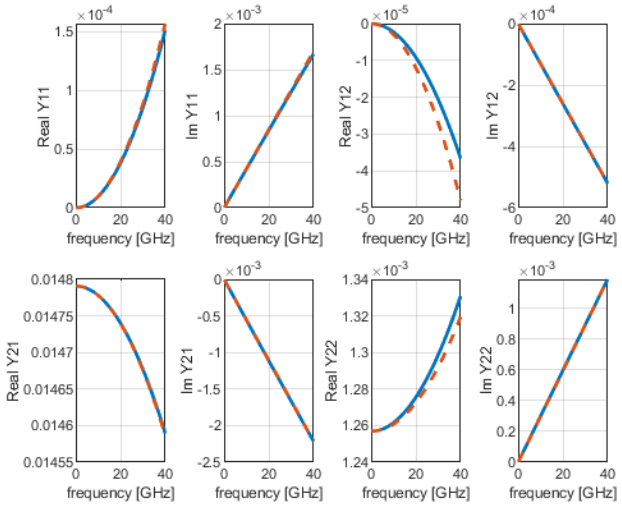
$$error = max\left(e_{ij,rms}\right) \tag{3.9}$$

where e_{ij} is the error matrix for one of the eight Y-parameters (four real and four imaginary parameters) and $e_{ij,rms}$ is the corresponding RMS error for each bias point over the frequency range. N is the number of frequency points. The maximum of the eight RMS error values is used to indicate the error at each bias point. Since the worst among the eight errors is selected, it is possible to track the Y-parameter for which the error is observed. This is helpful in case the parameters take unexpected values.

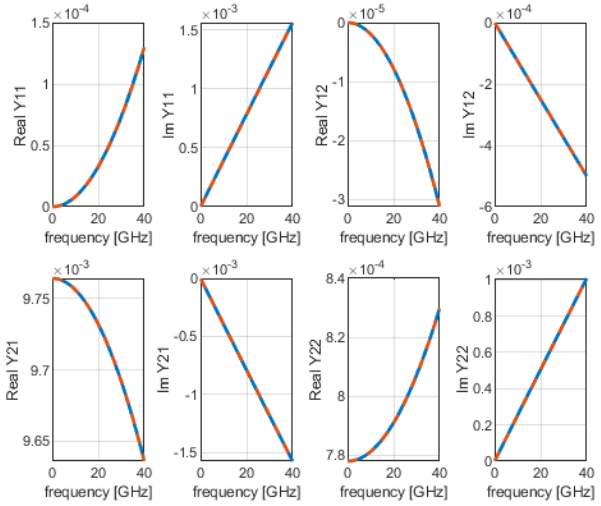
3.4.2. Study with simulated data from 22nm FDSOI technology

Based on the steps described above, the parameters for various bias points were extracted from S-parameter simulation data for a typical NMOS RF transistor(W = 500nm, L =18 nm, and NF = 16). The extraction steps were implemented in MATLAB. Figure 3.13 shows the PDK based Y-parameters vs model at $V_{GS} = 600m$ and $V_{DS} = 800m$. A small deviation is observed as the frequency increases. Figure 3.7b shows the comparison at $V_{GS} = 400m$ and $V_{DS} = 500m$ where the model data overlaps the PDK simulation.

Note that the real part of Y_{22} is deviating with increasing frequency. The gate capacitance, C_{bd} and C_m are extracted from the imaginary part of the Y-parameter which shows linear behavior with frequency and cannot cause this deviation. Thus, value extracted for R_b, R_g and C_{ms} might cause this. But note that it is cumbersome to point out which parameter is causing deviation as they are coupled to each other in the real part of the expressions. Although manually changing these parameters is possible to see which one is causing deviation.



(a) at $V_{GS} = 600mV$ and $V_{DS} = 800mV$.



(b) at $V_{GS} = 400mV$ and $V_{DS} = 500mV$

Figure 3.7: Comparison of y-parameters: PDK simulated (Solid lines) vs Small signal equivalent Model (dashed-lines)

3.4.3. Extracted Parameter variation with bias voltage

In the previous section, trial simulations at randomly selected voltage points showed that the simplified small-signal equivalent circuit is reasonably useful in capturing the small-signal behavior of the transistor. In order to generate a model for a FET, with a given dimension and operating temperature, the variation of its parameter set as a function of bias voltage needs to be captured. The variation should be such that a suitable function $f\left(V_{GS}, V_{DS}\right)$ can be formulated. This function shall be used by the simulator for the calculation of the device behavior. Additionally, it is an added advantage if the parameter variation can be qualitatively justified based on a physical understanding of the charge rearrangement that undergoes in the device.

Figure 3.8 shows the variation of the extracted parameters with V_{GS} for $V_{DS} = 800mV$. Compared to the rest of the parameters gate resistance stay relatively constant with bias voltage. Also, the value of R_b decreases as the bias voltage increases. This is similar to the result presented by Hung-Chi Han et. al. in [23] ($R_g \approx 40\Omega$ and $R_b \approx 10\Omega$).

This bulk resistance variation with inversion region in case of bulk CMOS was reported for saturation in [13]. Since we are using the same set of equations as for bulk CMOS without any technology-specific conditions in the extraction code, possibly R_B is varying due to that. Note that even for 40nm CMOS technology, which will be shown later, the Y-parameters show similar frequency dependency as for

22nm FDSOI.

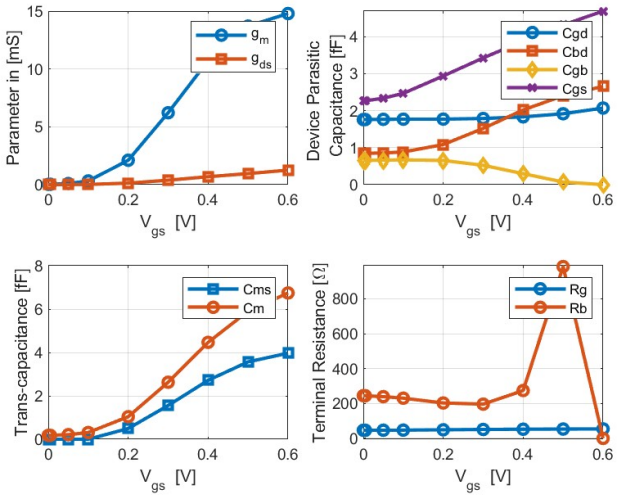


Figure 3.8: Extracted Parameter for various V_{GS} and $V_{DS} = 800mV$

The trans-conductance g_m and channel conductivity g_{ds} increase with gate voltage as expected. g_{ds} is relatively insensitive to gate voltage variation though. The total capacitance C_{gg} as seen in the gate terminal has three components C_{gs}, C_{gd}, and C_{gb}. C_{gs} is increasing with gate voltage as the device is moving into strong inversion region. C_{gd} is relatively constant as it models the overlap capacitance. C_{gb} tends to be constant at lower V_{GS} and with a decreasing trend as V_{GS} increases. As gate voltage increases, the FET channel goes more into inversion. This masks the gate to bulk capacitance which translates into a reduced C_{gb}.

Trans-capacitance C_{ms} and C_m showed an increasing trend with gate voltage. These two trans-capacitance parameters were included in the quasi-static model with first-order approximation to account for the frequency dependence of the trans-admittance. As the channel moves to strong inversion, an increase in the value of these two capacitors will take into account the inertia of charge in the channel. The effect of this becomes prominent with higher frequency of operation.

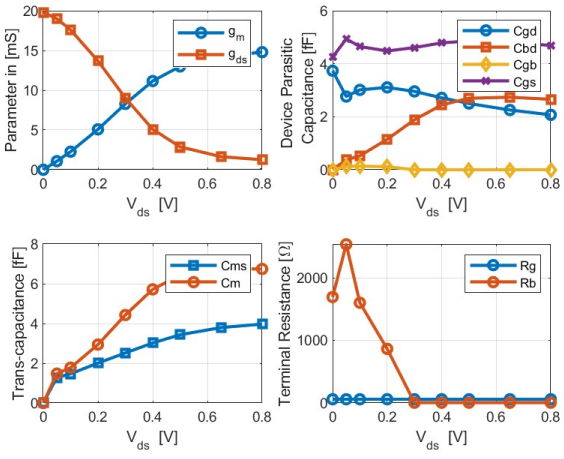


Figure 3.9: Extracted Parameter for various V_{DS} and $V_{GS} = 600mV$

Variation of parameters with V_{ds} for $V_{GS} = 600mV$ is presented in figure 3.9. As expected, g_m is increasing while channel conductivity is decreasing with increasing drain voltage. Note that the gate resistance is constant as before. The bulk resistance takes very small values as the bias voltage increases. Gate-drain capacitance C_{gd} is decreasing with drain voltage while C_{gs} remains approximately

constant. The bulk-drain capacitance increases with V_{DS} but saturates eventually. As the drain voltage increases the channel becomes narrow near the drain region. A small increase in the capacitance can be due to that. The variation of the parameters is similar to the analysis presented in [35] (Chapter-9, pages 453-460 and Chapter-8, figure 8.8).

Since the final goal is to build a model card that can be used in a wide range of bias points, the behavior of the small-signal parameters in various regions of operation should also be understood. The usual assumption of small-signal equivalent circuit modeling for RF CMOS transistors is that the device shall be biased in the saturation region. The choice of bias point is application dependent and up to the circuit designer. Even if the model parameters lose physical meaning at a certain region of operation, the accuracy of device behavior is prioritized. The purpose of analyzing the model response for a wide range of bias points is to identify the operating region limitation of the small-signal equivalent circuit.

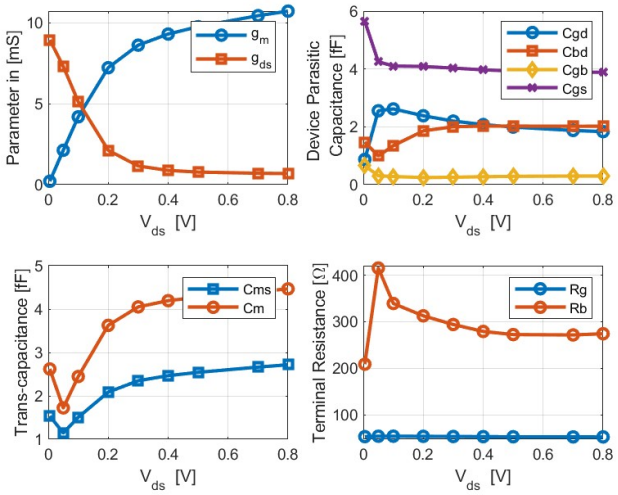


Figure 3.10: Extracted Parameter for various V_{DS} and $V_{GS} = 400mV$

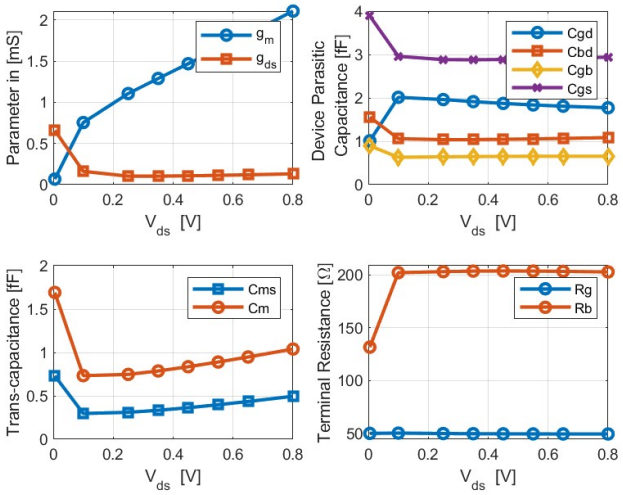


Figure 3.11: Extracted Parameter for various V_{DS} and $V_{GS} = 200mV$

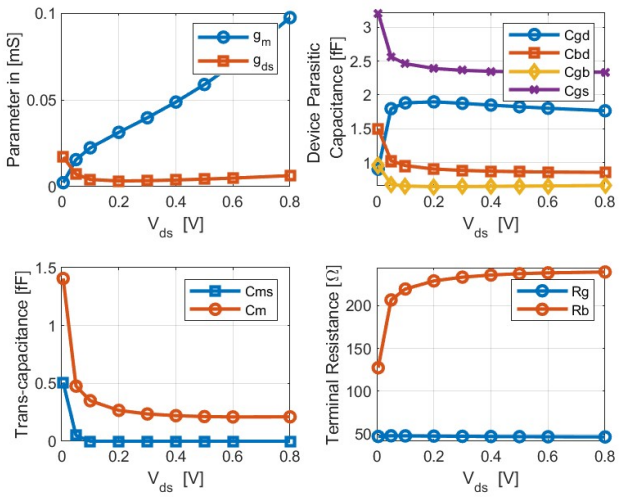


Figure 3.12: Extracted Parameter for various V_{DS} and $V_{GS} = 50mV$

Figures 3.10 to 3.12 show the extracted parameters vs V_{DS} for different V_{GS}. The gate capacitance decreases with V_{GS} while staying approximately constant with V_{DS}. Note how the variation in bulk resistance with V_{DS} decreases as the gate-source voltage decreases,i.e., device channel moves towards weak-inversion region. As the device is moving towards weak inversion the trans-capacitance effect reduces. Also, Bulk to drain capacitance changes its trend and starts decreasing with increasing V_{ds} and eventually reaches a saturation region. It is worth noting that the variation of bulk resistance rather non-uniform with V_{gs}.

Figure 3.13 shows the error between the Y-parameters simulated from PDK and the small signal model at various bias voltages. The error is calculated using equation 3.9. Although the error in general is small but from 3.13 we can expect some deviation at higher values of V_{GS}.

One can use such a method to define errors in order to automate the parameter optimization using various non-linear curve fitting functions. But defining error in Y-parameter fitting is rather non-intuitive as a standalone metric from an analysis point of view. Thus other common RF performance metrics like f_t and f_{max} may be more useful.

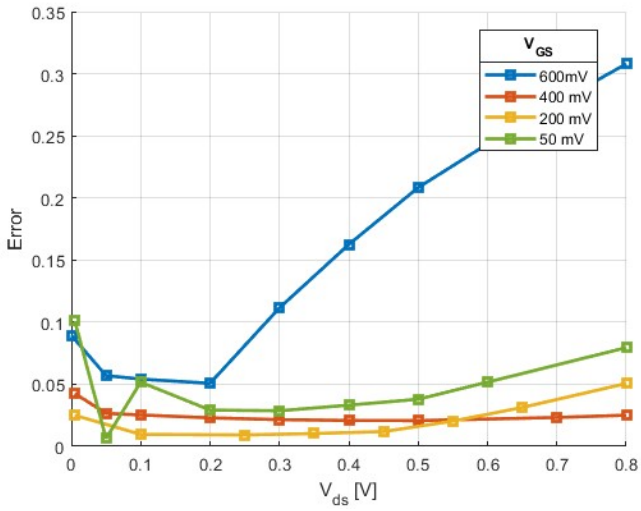


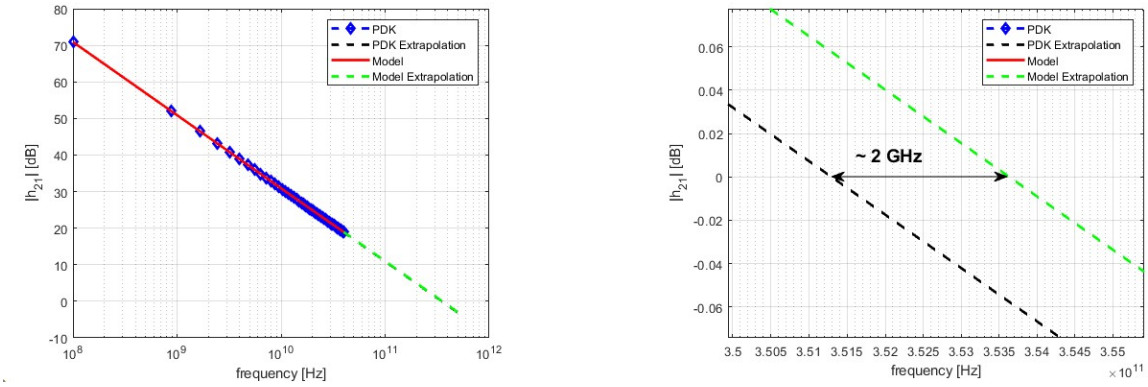
Figure 3.13: Error between PDK simulation data and model data for various V_{DS} and V_{GS}. For V_{GS} 600mV, real of Y_{12} and real of Y_{21} is deviating (see figure)

3.4.4. f_t and f_{max} comparison

The error plots in the previous section give an indication of the deviation of the model from the PDK. But note that it was based on RMS error of four Y-parameters each having a real and imaginary part. It cannot be used to compare two models for accuracy. Even though no widely accepted metric is available for comparison we can use the RF performance metric to test how accurately these are calculated by the model.

For f_t determination at any given bias point, current gain $|H_{21}|$ expression is equated to 1 and solved for frequency with the parameters known from earlier extraction process. Figure 3.14a show the current gain calculated from PDK and model at $V_{GS} = 600mV$ and $V_{DS} = 800mV$. Figure 3.14 shows the zoom-in near frequency where gain is approaching unity. The PDK data has been extrapolated from 40GHz at -20dB/decade. Transit frequency values determined by the model(∼351.5GHz) and PDK(∼353.5GHz) show deviation of around 13 GHz. Since measurement till such high frequency is difficult, extrapolation is a widely accepted method to determine f_t.

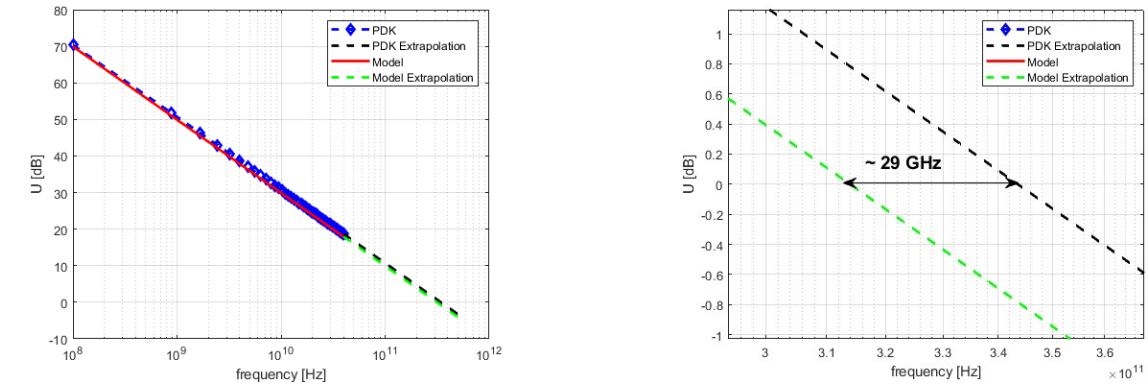
f_{max} corresponds to the unity value of Mason's unilateral power gain (U). Using a similar methodology as in f_t, U as extrapolated from the model simulation is compared to values determined by extrapolation of PDK data. Model calculates f_{max} of ∼ 315GHz while PDK extrapolation gives ∼ 344GHz. A 29 GHz deviation is observed.



(a) Current gain $|h_{21}|$ vs frequency

(b) Zoom-in near unity gain frequency

Figure 3.14: Comparison of current gain and f_t at $V_{GS} = 600mV$ and $V_{DS} = 800mV$



(a) Unilateral power gain U vs frequency

(b) Zoom-in near unity gain frequency

Figure 3.15: Comparison of unilateral gain and f_{max} at $V_{GS} = 600mV$ and $V_{DS} = 800mV$

The current gain and power gain plots show very good agreement at lower frequencies. Though deviation is observed between the f_t (∼ 3.6% error) and f_{max} (∼ 12.46% error) so determined.

3.5. Usability with TSMC 40nm bulk CMOS technology

While the small-signal model and corresponding extraction method were able to replicate a typical low threshold voltage RF transistor in 22nm FDSOI technology with a fair level of accuracy, here we test if the same can be used for transistors from the TSMC 40nm technology. We follow the same procedure for analysis as before. The admittance parameters from the PDK simulation are used to extract the model parameters. The model is simulated and compared with the PDK simulation.

Figure 3.18 shows the variation of the parameters with V_{ds} at different V_{gs}. g_m increases with V_{ds} while g_{ds} shows gradual decrease.For any given value of V_{gs}, the capacitance seen at the gate terminals stays almost constant with V_{ds} while the C_{bd} shows stronger dependency on V_{ds}. For $V_{gs} = 500mV$ and above, C_{bd} will increase with V_{ds} and saturate. But for lower V_{gs} the same will show a decaying trend. While gate resistance remains constant, the bulk resistance decreases with increasing V_{ds}. The bulk resistance also shows a decreasing trend with V_{gs}. Trans-capacitance C_{ms} and C_m are increased with drain voltage and saturate at a strong inversion region. At weaker inversion regions the trans-capacitance effect is greatly reduced.

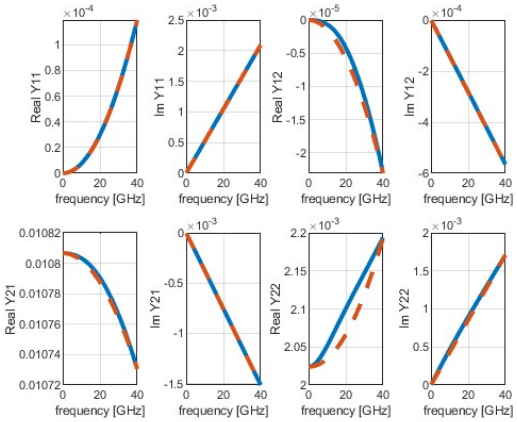


Figure 3.16: Comparison of Y-parameter: $V_{DS} = 1.1V$ and $V_{GS} = 1.1V$. (Solid blue line is PDK simulated and dashed orange line is from model.)

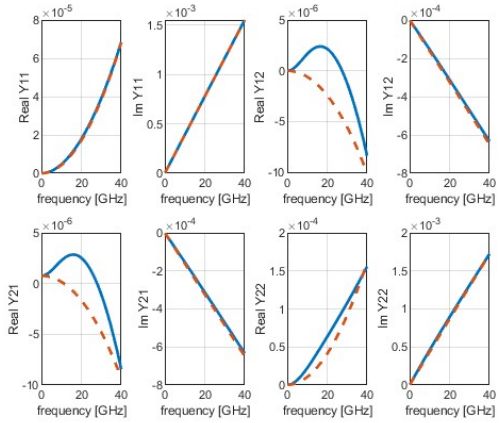
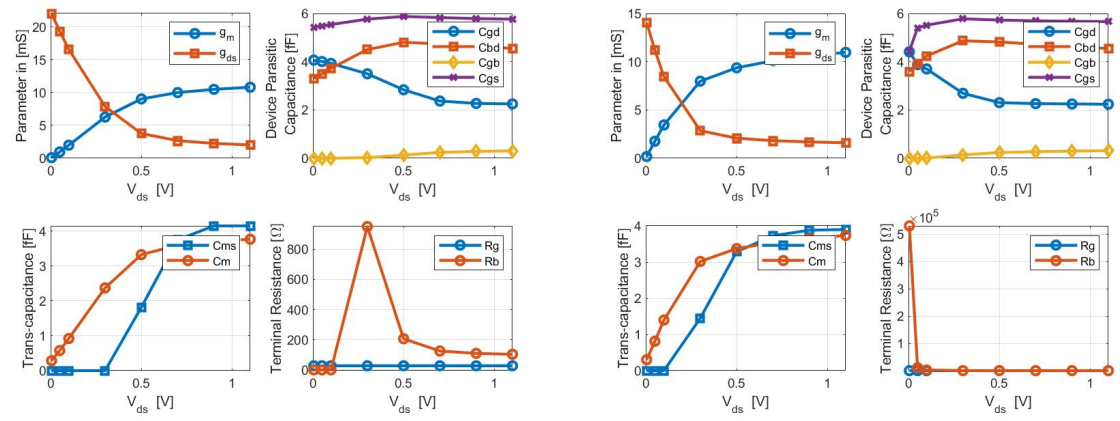
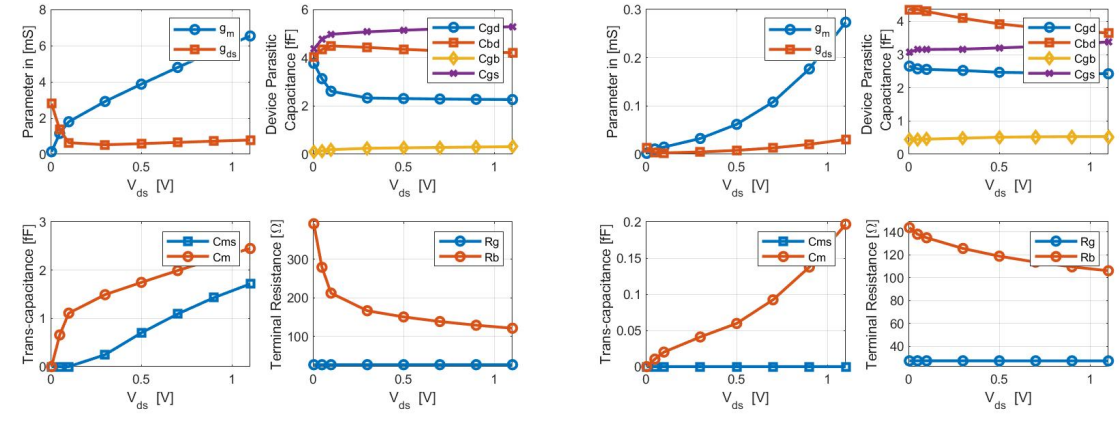


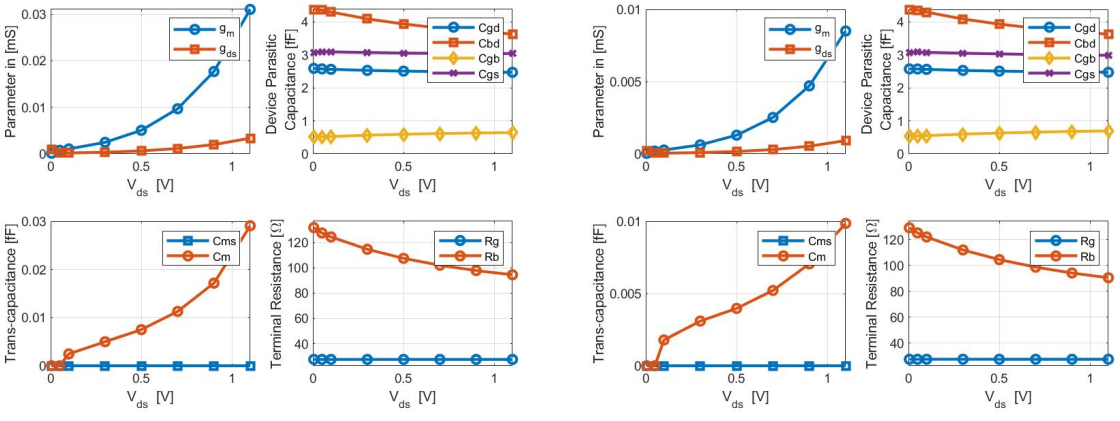
Figure 3.17: Comparison of Y-parameter: $V_{DS} = 50mV$ and $V_{GS} = 100mV$. (Solid blue line is PDK simulated and dashed orange line is from model.)



(a) $V_{GS} = 1100mV$ **(b)** $V_{GS} = 800mV$



(c) $V_{GS} = 500mV$ **(d)** $V_{GS} = 200mV$



(e) $V_{GS} = 100mV$ **(f)** $V_{GS} = 50mV$

Figure 3.18: Extracted Parameter for various V_{DS} and V_{GS}

The error between the model and the PDK simulation is shown in figure 3.19. The error values we get in this case are rather high and the reason can be justified by the plot in 3.16. The real of Y_{22} is modeled quadratic dependency on frequency. But the PDK shows higher-order behavior. Thus the model is limited in this respect.

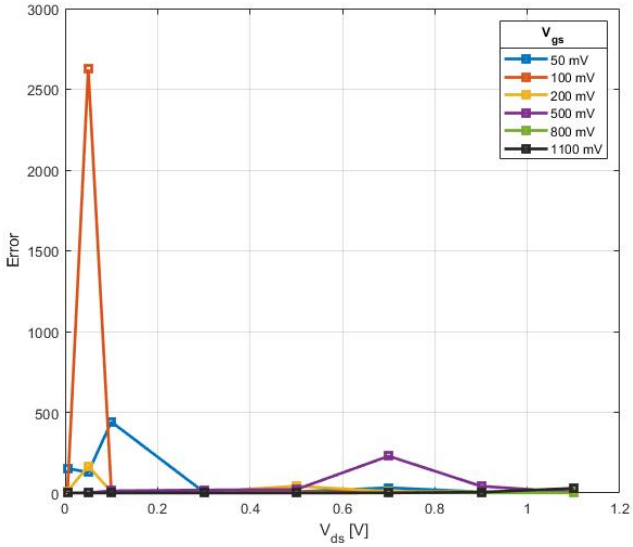
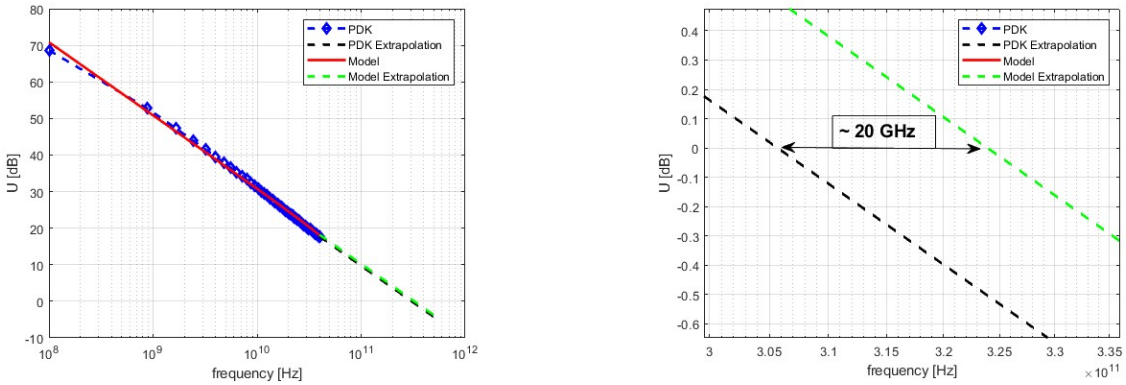


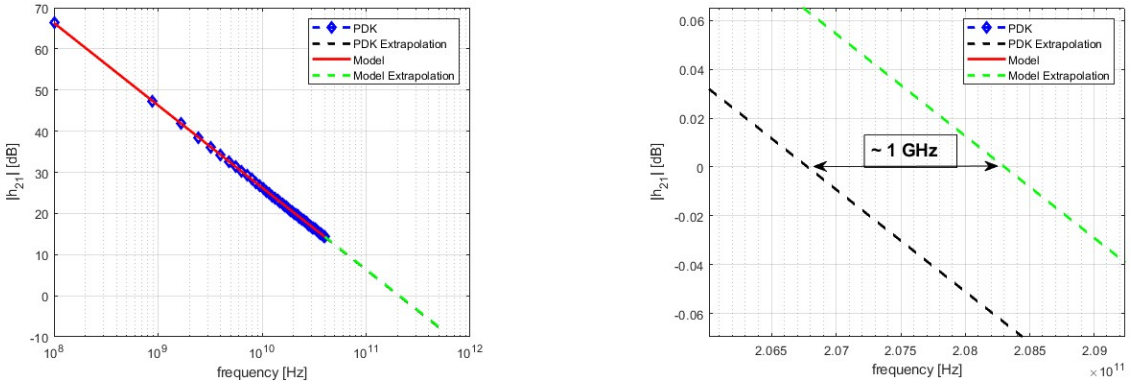
Figure 3.19: Error between Y-parameter from PDK simulation data and model data for various V_{DS} and V_{GS}. The error goes very high in weak inversion because the dependency of real parts of Y_{12} and Y_{21} on frequency is not quadratic. See figure 3.17



(a) Unilateral power gain U vs frequency

(b) Zoom-in near unity gain frequency

Figure 3.20: Comparison of unilateral gain and f_{max} at $V_{GS} = 1.1V$ and $V_{DS} = 1.1V$



(a) Current gain $|h_{21}|$ vs frequency

(b) Zoom-in near unity gain frequency

Figure 3.21: Comparison of current gain and f_t at $V_{GS} = 1.1V$ and $V_{DS} = 1.1V$

3.6. Merits and Demerits of the small-signal equivalent circuit

Due to the simplifications the parameter extraction process for the model parameters are determined in a straight forward way. The equivalent circuit was able to replicate the Y-parameter response of the device with reasonable accuracy compared to PDK based simulation. As can be seen in figure 3.16, some deviation is observed in general as the model is able to capture only quadratic dependency on frequency of the Y-parameter. For the case of TSMC 40nm bulk CMOS technology, Y_{22} as simulated from the PDK does not show a quadratic dependency on frequency.

The non-quasi-static nature of FET becomes more prominent as the frequency of operation increases. The Y-parameter comparison till 40GHz does not show this effect as the values calculated by the model is very close to that from PDK. While the current gain and unilateral gain plots overlap at frequencies below 100GHz; deviation can be observed beyond that. As a result the f_t and f_{max} calculation show deviation.

3.7. Implementation in Verilog-AMS

The extracted parameters are used to train the ANN for each of the parameters and a model is generated. This is done with the ICCAP ANN tool. At the end of ANN training, ICCAP saves the generated function in the assigned directory with the ".inc" extension. This file which contains the function definition is usable in the Verilog-AMS implementation of the small signal model we used to extract the model parameters, by defining the location of the file using the "'include" statement.

For any pair of gate and drain terminal voltages, the function will evaluate the small signal model parameters and the gate and drain DC current. These two currents can be implemented in Verilog-A using the contribution operator ("<+") as shown below.

$I(g,s) <+ I_g;$
$I(d,s) <+ I_d;$

For the implementation of the capacitors in the small signal model, the corresponding charge is calculated first. The time derivative operator is used to determine the current through it.

$Q_{ij} = C * V(i,j);$
$I(i,j) <+ ddt(Q_{ij});$

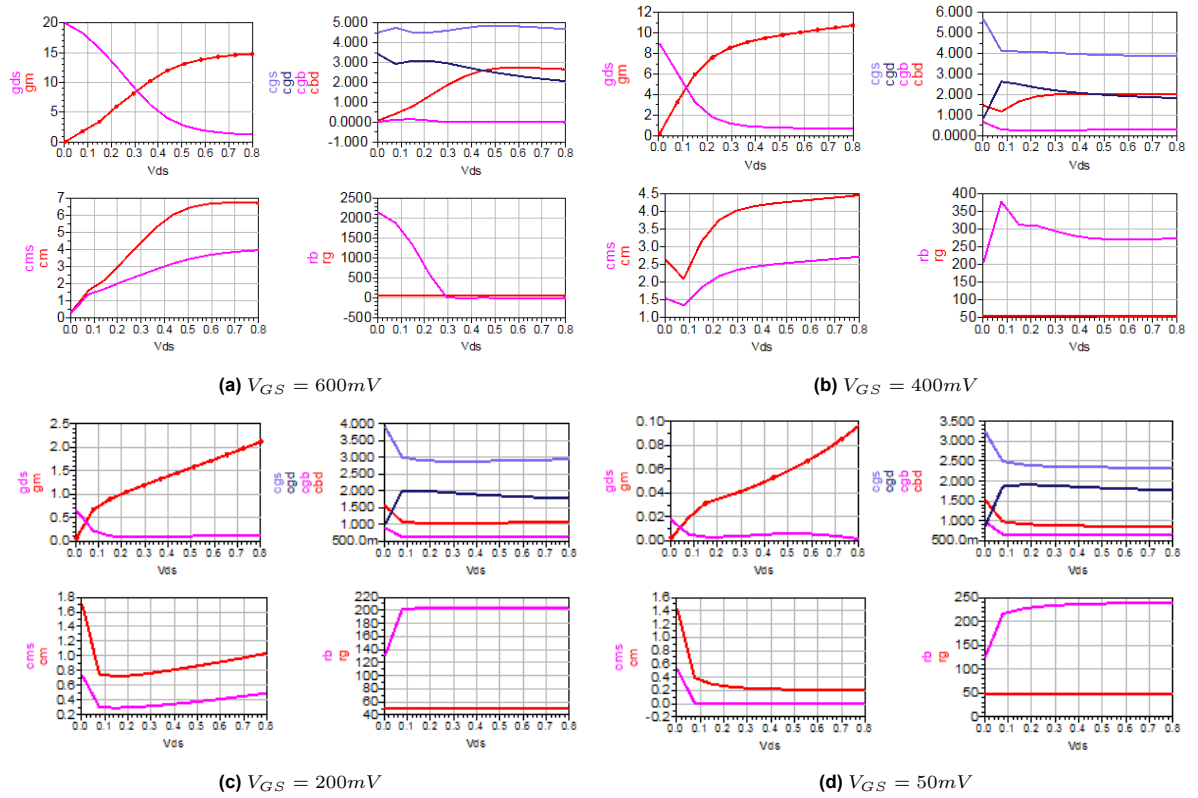
Note that C_m and C_{ms} are an artifact of the non-quasi-static nature of the transistor channel. As was shown in the equivalent circuit they act in parallel to the trans-admittance as a reactive voltage-controlled current source. Their implementation is shown below. The voltage across nodes p and n contributes a reactive current between nodes i and j.
$Q_{ij} = C * V(p,n)$
$I(i,j) <+ ddt(Q_{ij})$

Similarly, the trans-conductance can be implemented as
$I(i,j) <+ g_m * V(p,n)$

Thus, the small signal model of the transistor can be implemented. This can then be used for AC simulations.

3.7.1. Simulation Results

Figure 3.22 shows the values of the parameters calculated by the ANN generated function for the case of the NMOS transistor from GF 22nm FDSOI discussed in section 3.4.3. The ANN generated model parameters are in good agreement with the values extracted as shown earlier.



(a) $V_{GS} = 600mV$ **(b)** $V_{GS} = 400mV$

(c) $V_{GS} = 200mV$ **(d)** $V_{GS} = 50mV$

Figure 3.22: Model Parameters calculated from ANN generated function for various V_{DS} and V_{GS}

Figure 3.23 shows the simulated Y-parameters at $V_{DS} = 800mV$ and $V_{GS} = 600mV$. Note that the PDK-simulated data and ANN-generated model simulations have deviations from each other. But what needs to be understood here is that the model generation flow till now has no feedback loop. The ANN size used to build the model equation was determined heuristically. The model has four input variables, V_{gs}, V_{ds}, width W and length L. Temperature is not considered here. ANN should be able to map the variation introduced by four variables on the ten model parameters. A logical point to start will be to select 4 hidden layers and each layer shall have 10 neurons. This should be sufficient to map the variation of the parameters with input variables. But we also note that some of the model parameters do not show a gradual rise or fall trend. To accommodate such variations, 12 neurons per hidden layer is selected although this is the best guess number. Hence, it is justified that the final model equation generated will invariably introduce some error. Although note that the deviation is not drastic. Apart from the real of Y_{22}, which is a small figure in itself, other parameters stay close to each other. Using the error metric defined previously, the error between PDK and model simulation is 16.14%, which is significant.

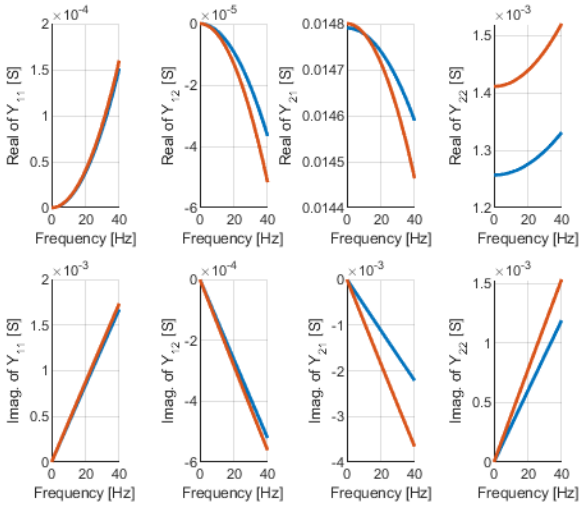
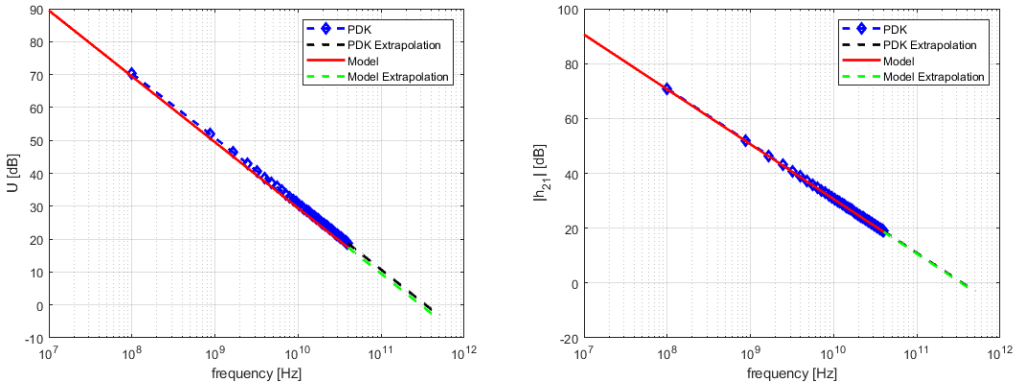


Figure 3.23: Comparison of Y-parameter: $V_{DS} = 800mV$ and $V_{GS} = 600mV$. (Solid blue line is PDK simulated and dashed orange line is from ANN model.)



(a) Unilateral Power Gain vs Frequency

(b) Current Gain vs Frequency

Figure 3.24: Unilateral Power gain and Current Gain simulation of the ANN model at $V_{DS} = 800mV$ and $V_{GS} = 600mV$.

Figure 3.24 shows the device's unilateral gain and current gain for the same DC operating point. The extrapolation lines in 3.24, show $f_t = 341GHz$ and $f_{max} = 296GHz$ from the ANN model simulations.

This analysis shows that the f_{max} and f_t values determined by the model stay in close proximity to PDK simulations. Although the model is reliable at low frequency region, accuracy deteriorates at higher frequencies.

3.8. Summary

In this chapter, a simplified small-signal equivalent circuit was discussed and its feasibility for modeling RF CMOS transistors from two different technologies was investigated. The model generation flow using ICCAP's ANN model generation tool was presented. The parameters at various bias points were analytically determined. Without the utilization of any optimization algorithm, the generated model was reasonably close to the PDK data. The shortcomings of the proposed modeling approach was also presented.

It is worth mentioning here that, the input data set used here is ideal without any outliers. But in practical measurement data, one should expect scattered data points. The PDK Y-parameter is very smooth. But the measured device Y-parameter will not be so and as such it will be necessary to use some functions that will help to determine the best fitting curves. Thankfully, quite a few of these functions can be found in MATLAB.

The use of ANN helps to generate model functions which would be otherwise rigorous work to do. Especially at cryogenic temperatures where temperature scaling laws, used for commercial temperature range, lead to significant deviation from measurement [5], significant effort is required to build a physical model for MOS transistors. This is an active area of research. From a circuit design perspective though, a model that can replicate the electrical behavior of the device accurately is desired. This can be done with ANN-generated models.

4

Probing a device RF characterisation fixture

This chapter is dedicated to reviewing the structure of characterization chip. Followed by that we also review a calibration technique and a de-embedding technique. The use of EM simulation to get the standard definition is also mentioned briefly. Finally some considerations for cryogenic measurement setup is also mentioned.

4.1. RF Characterisation Chip Structure

For RF characterization of MOS transistors using S-parameters, the device is embedded between two transmission lines with the gate and drain forming the two signal ports. The transmission line structures connect the gate and drain of the MOSFET to the signal pads at the top metal layer where the RF probe tips make contact. The source and bulk terminals are tied together and connected to the ground net of the TL structure.

Figure 4.1 shows a representative diagram of the RF characterization setup. Assuming the VNA is ideal till its ports, the S-parameter measurement of the device under test (a.k.a DUT) will require the removal of the effect of the coaxial cables and the probe itself by the calibration procedure. In this work, the short-open-load-reciprocal calibration procedure is used. The on-wafer calibration structure has a short, open, thru, and 50 Ω load placed in place of the DUT. These form the known standards. The characteristics of these four known structures are known from electromagnetic simulations in the desired frequency range. The reciprocal standard is a low resistance thru structure whose accurate response is not required for a short-open-load-reciprocal calibration procedure. Using this calibration the reference plane of S-parameter measurement is shifted from the VNA ports closer to the device embedded in the characterization fixture.

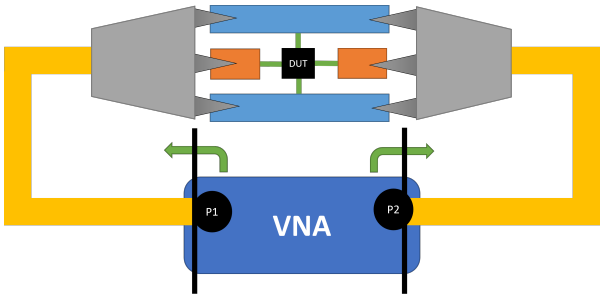


Figure 4.1: Characterisation setup representative diagram

Figure 4.2 shows a three-dimensional view of a typical characterization structure under consideration. Note that the actual device to be characterized is placed at the bottom and is connected to the pads using through-silicon-vias.

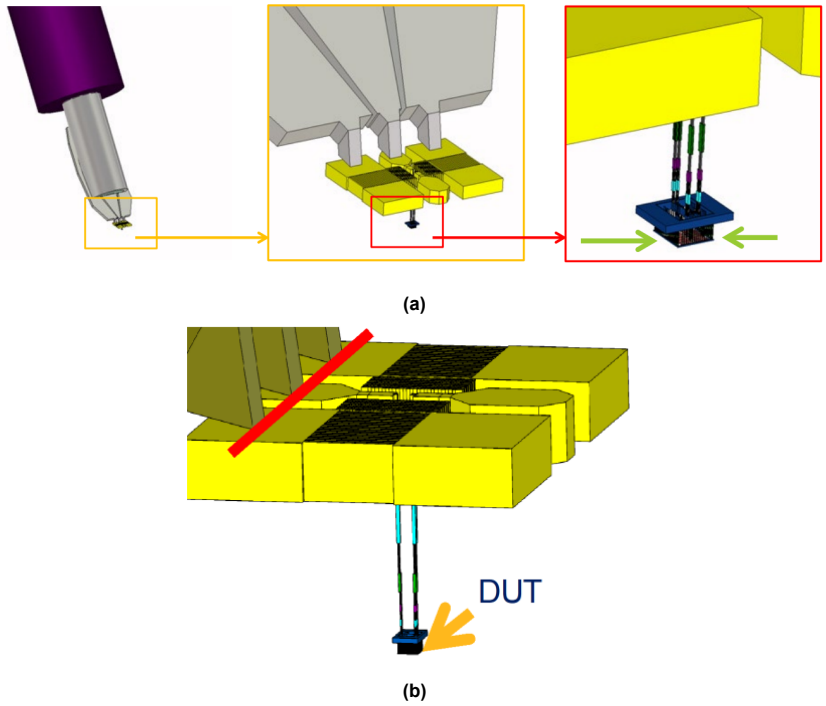


Figure 4.2: 3-dimensional view of typical characterization structure [32]

While calibration will move the measurement reference plane from VNA ports to the probe tip (see the red line in 4.2), from a device modeling perspective the s-parameter response of the device at the lowest metal layers (usually named M1) is desired. For the frequency range under consideration, the additional metal layers add capacitance between the terminals and the terminal to the ground. Given the characterization structure, it becomes necessary to remove the effect of the pad structure. In this work, two-step open-short de-embedding is used. This removes the parasitic capacitance and resistance of the pad to M1 layer structures. In the sections to follow a short review of the SOLR calibration is presented. Interested readers can refer to various literature cited hereafter for an in-depth description of the same.

4.1.1. Brief review of S-parameters

In order to characterize the behavior of two-port networks, both the transfer and impedance functions must be measured. The available tools are Z, Y, h, ABCD, and scattering parameters. In order to determine these parameters the two ports of the network are measured at different configurations of open and short-circuit tests. While this is a straightforward procedure, at microwave frequencies it becomes difficult to achieve the ”short and open circuit conditions with respect to ac signals” over a wide frequency range. Note that the cable from the VNA port to RF probes is to be considered a two-port network with distributed elements. Thus scattering parameters (S-parameters), which are defined in terms of traveling waves and capture the complete behavior of two-port networks, are used. It is easy to use in analysis and flow graph theory is directly applicable too [21].

Consider the two-port network shown in figure 4.3. a_i and b_i are the incident and reflected waves respectively. Z_{01} and Z_{02} are the characteristic impedance of the corresponding transmission lines.

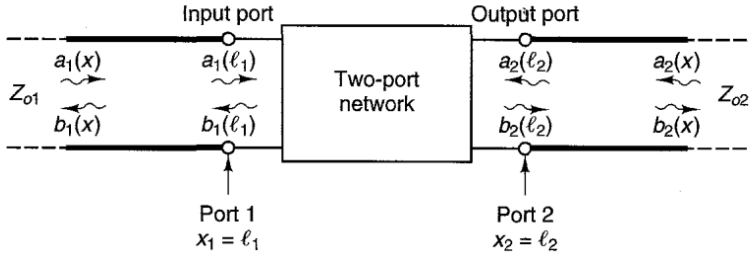


Figure 4.3: Incident and reflected waves in a two port network [21]

We can define the scattering parameters of the two ports of the network as,

$$b_1(l_1) = S_{11}a_1(l_1) + S_{12}a_2(l_2) \tag{4.1}$$

$$b_2(l_2) = S_{21}a_1(l_1) + S_{22}a_2(l_2) \tag{4.2}$$

The term $S_{11a1}(l_1)$ represents the contribution of the incident wave $a_1(l_1)$ to the reflected wave $b_1(l_1)$. Similarly, the rest of the parameters can be defined. S-parameters as such can be visualized as representing the reflection or transmission coefficients at the two terminals.

Since we want to characterize CMOS transistors, the two-port network will be representative of it. This is a black-box approach to characterize the transistors. In figure 4.3, the transmission lines, are cascade-connected to the two-port network (the DUT). In order to analyze such a configuration, the chain scattering parameters, also called scattering transfer parameters (T-parameters) is often useful. It is defined as below.

$$a_1(l_1) = T_{11}b_2(l_2) + T_{12}a_2(l_2) \tag{4.3}$$

$$b_1(l_1) = T_{21}b_2(l_2) + T_{22}a_2(l_2) \tag{4.4}$$

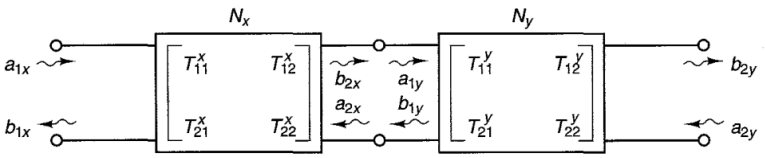


Figure 4.4: Incident and reflected waves in a two port network[21]

For the networks N_x and N_y in figure 4.4, the incident and reflected waves at the port 1 of N_x can be represented in terms of the waves at port 2 of N_y as shown in

$$\begin{bmatrix} a_{1x} \\ b_{1x} \end{bmatrix} = \begin{bmatrix} T_{11}^x & T_{12}^x \\ T_{21}^x & T_{22}^x \end{bmatrix} \begin{bmatrix} T_{11}^y & T_{12}^y \\ T_{21}^y & T_{22}^y \end{bmatrix} \begin{bmatrix} b_{2y} \\ a_{2y} \end{bmatrix} \tag{4.5}$$

The T-parameters can be defined in terms of s-parameters.

$$\begin{bmatrix} T_{11} & T_{12} \\ T_{21} & T_{22} \end{bmatrix} = \begin{bmatrix} \frac{1}{S_{21}} & -\frac{S_{22}}{S_{21}} \\ \frac{S_{11}}{S_{21}} & S_{12} - \frac{S_{11}S_{22}}{S_{21}} \end{bmatrix} \tag{4.6}$$

In the section to follow, this matrix representation of network behaviour will be used to review the calibration procedure. Although it might seem cumbersome to use S-parameters, the direct translation into matrix form is very useful in data processing using tools like MATLAB. The S-parameter matrices under consideration are 2 x 2 square matrices at large number of frequency points and a wide span of voltage bias points. Hence implementation in the form of three dimensional array improves the speed of operation in MATLAB.

4.2. Review of S-O-L-R calibration

4.2.1. Sources of error in S-parameter measurement

In order to understand the need of calibration the sources of error in s-parameter measurement needs to be identified first. A generic four channel vector network analyzer block diagram is shown in figure 4.5[30]. Figure 4.6 shows a signal flow graph that shows all the possible signal flow paths when the network analyzer is excited at port 1 and port 2 is receiver end. When VNA is excited at port 2 with port 1 match terminated, the signal flow graph will be similar but in reverse order.

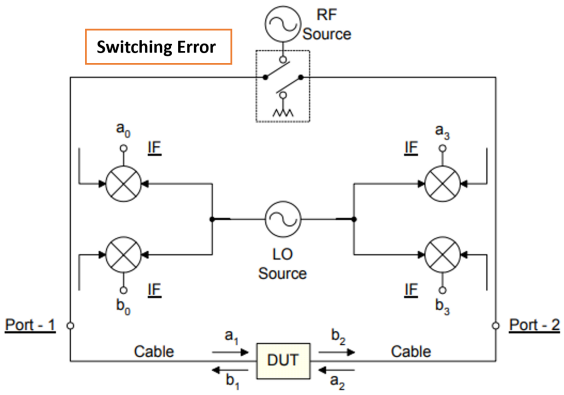


Figure 4.5: Generic block diagram of a vector network analyzer [30]

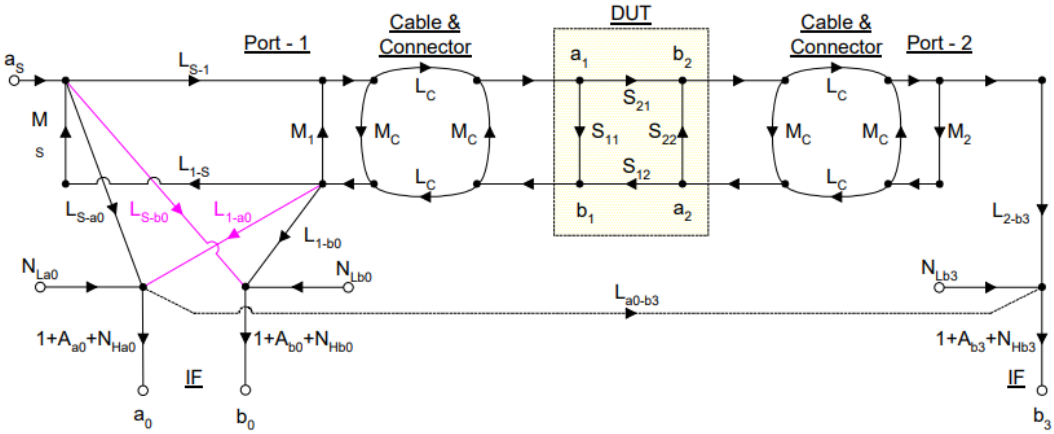


Figure 4.6: Complete Signal Flow Graph describing all possible signal flow paths for forward direction operation of VNA[30]

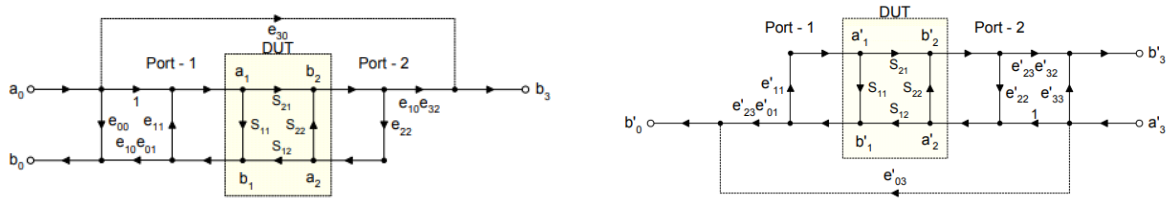
Below is the description of the error terms mentioned in figure 4.6.

1. a_1 = Incident Signal at Port-1
2. b_1 = Reflected Signal at Port-1
3. a_2 = Incident Signal at Port-2
4. b_2 = Transmitted signal at Port-2
5. a_S = Source Port
6. a_0 = Measured Incident Port
7. b_0 = Measured Reflected Port
8. b_3 = Measured Transmitted Port
9. L_{S-1} = Loss from Source to Port-1
10. L_{1-S} = Loss from Port-1 to Source
11. L_{S-a0} = Loss from Source to a_0

12. L_{S-b0} = Loss from Source to b_0 (Directivity)
13. L_{1-a0} = Loss from Port-1 to a_0 (Directivity)
14. L_{1-b0} = Loss from Port-1 to b_0
15. L_{2-b3} = Loss from Port-2 to b_3
16. L_{a0-b3} = Loss from a_0 to b_3 (Leakage)
17. L_C = Loss of Cables
18. M_1 = Match at Port-1
19. M_2 = Match at Port-2
20. M_S = Match of Source
21. M_C = Match of Cables
22. N_L, N_H = Noise at the node
23. A = Dynamic Accuracy at the node (Linearity)

4.2.2. Error model for linear calibration

The signal flow graph in 4.6 is detailed and includes the system non-linearity and noise. A simple linear calibration algorithm like SOLR cannot correct for noise and non-linearity. Since the IF bandwidth is kept high the noise will be reduced and can be neglected in most cases. System linearity is assumed for this work. A simplified version of this error model is shown in figure 4.7.



(a) Forward direction **(b)** Reverse direction

Figure 4.7: Reduced Signal Flow Graph ignoring noise and system non-linearity[30]

In figure 4.7, the power received at port 1 of DUT is considered as ideal and all the error terms are normalised accordingly. There are 6 error terms to be calculated in each direction. The same is true with reverse model. This is also referred to as 12 term error model.
e_{00} is the directivity and is the ratio of loss from source to b_0 and loss from source to a_0. $e_{10}e_{01}$ is the reflection tracking. $e_{10}e_{32}$ is the transmission tracking. e_{11} and e_{22} denotes port match at ports 1 and 2 respectively. e_{30} is the leakage from port 1 to port 2. The same is define for reverse direction.
Once these error terms are known, the actual S-parameters of the DUT is determined by solving the following four equations.

$$S_{11M} = \frac{b_0}{a_0} = e_{00} + (e_{10}e_{01})\frac{S_{11} - e_{22}\Delta_s}{1 - e_{11}S_{11} - e_{22}S_{22} + e_{11}e_{22}\Delta_s} \tag{4.7}$$

$$S_{21M} = \frac{b_3}{a_0} = e_{30} + (e_{10}e_{32})\frac{S_{21}}{1 - e_{11}S_{11} - e_{22}S_{22} + e_{11}e_{22}\Delta_s} \tag{4.8}$$

$$S_{12M} = \frac{b'_0}{a'_0} = e'_{03} + (e'_{01}e'_{23})\frac{S_{12}}{1 - e'_{11}S_{11} - e'_{22}S_{22} + e'_{11}e'_{22}\Delta_s} \tag{4.9}$$

$$S_{22M} = \frac{b'_3}{a'_3} = e'_{33} + (e'_{32}e'_{23})\frac{S_{22} - e'_{11}\Delta_s}{1 - e'_{11}S_{11} - e'_{22}S_{22} + e'_{11}e'_{22}\Delta_s} \tag{4.10}$$

where $\Delta_s = S_{11}S_{22} - S_{12}S_{21}$.

4.2.3. 8 term Error Model

In order to determine the error terms described earlier at least one of the calibration standards should have complete two port description. Usually this is a thru standard although obtaining accurate knowledge of it is often rigorous if not impossible [20]. In such a situation the short, open and load standards can be used to carry out one port calibration. Careful design of the characterisation structure makes the leakage terms e_{30} and e'_{03} between the ports negligibly small to be considered in calculations. In such a case the error model can be define as shown in figure 4.8 [20].

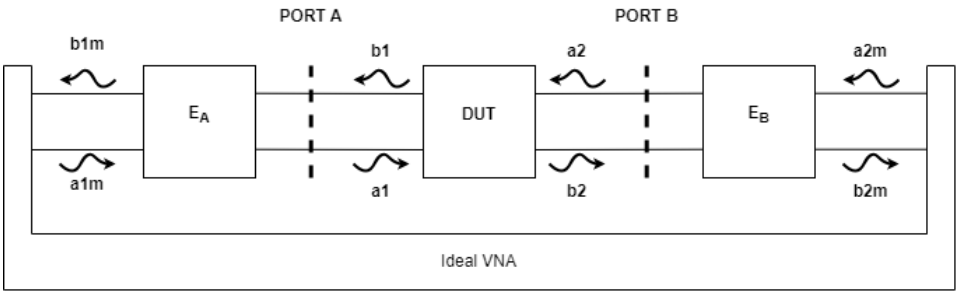


Figure 4.8: Error Model used for SOLR calibration [20]

This error model has only eight error terms represented by the error-boxes E_A and E_B. In order to determine the error terms single port calibration procedure on each of the ports using short, open and load standards is performed. The thru standard is a passive two port network such that $S_{12} = S_{21}$ using this property the final calibration step is performed. Once all the error terms are known the actual device s-parameters can be calculated from the measured S-parameters. As proposed in [20], the error boxes are defined with scattering matrices $E_A = \begin{bmatrix} e_A^{00} & e_A^{01} \\ e_A^{10} & e_A^{11} \end{bmatrix}$ and $E_B = \begin{bmatrix} e_B^{00} & e_B^{01} \\ e_B^{10} & e_B^{11} \end{bmatrix}$ with the terms having same definition as defined earlier. The calibration procedure requires the knowledge of reciprocity and the phase of the thru standard. The reflection coefficient measured at the ideal VNA port on port A side can be defined as

$$\Gamma_{mA} = \frac{e_A^{00} - \Gamma_A \cdot \Delta_A}{1 - \Gamma_A \cdot e_A^{11}}$$

where $\Delta_A = e_A^{00} e_A^{11} - e_A^{10} e_A^{01}$. By measuring three known standards (open, short and load) where Γ_A is known we can determine e_A^{00}, e_A^{11} and Δ_A. The same is repeated for port B. Note that the error-boxes and the device are cascaded to each other. In such a case transmission matrix representation will allow to write

$$T_m = \alpha \cdot T_A \cdot T_{DUT} \cdot T_B^{-1} \tag{4.11}$$

where T_m is the measured data, T_{DUT} is the known device two port response, $T_{A,B}$ are error matrices and $\alpha = e_A^{01}/e_B^{01}$, is an unknown. α is determined from the knowledge of reciprocity and a rough knowledge of the phase of S_{21}. Note that α is there in the above expression due to the transmission matrix representation of the error matrices. These calculation steps are not shown here and can be referred to in [20]. It is required in order to calculate all the error terms in E_A and E_B. Due to reciprocity, the determinant of the transmission matrix of a thru standard will be unity. Thus from equation 4.11 we can write for the through measurement,

$$det(T_m) = \alpha^2 \cdot det(T_A) \cdot det(T_B^{-1}) \tag{4.12}$$

From the above equation, two values, α with opposite phase is determined. At this stage, from one-port calibration T_A and T_B matrices are known. From the measured T_m matrix of thru standard, we can rewrite equation 4.11 as

$$X = \alpha T_{DUT} = T_A^{-1} \cdot T_m^{thru} T_B \tag{4.13}$$

By definition, $X(2,2) = \frac{\alpha}{S_{21}^{thru}}$. Using this relation the correct sign of α can be determined. A rough knowledge of the phase of S_{21}^{thru} is enough to determine the correct value of alpha.

4.2.4. Switch term correction

Till now the error model so presented neglected the leakage of the non-ideal electronic switch (see figure 4.5) [30] used to switch between VNA RF power source and Z_0 termination inside it. This is corrected by taking into account the mismatch due to the switches which can be measured directly in the case of a four-sampler VNA. Note that only port mismatch was considered and anything internal to the VNA was assumed to be ideal (see figure 4.8).

A thru structure transmits the power incident from port 1 of VNA which will be received at port 2 of VNA, with some loss due to the non-ideal thru. The switch terminates port 2 to Z_0. Due to the leakage in the switch, some of the power received at port 2 will be reflected back. Thus, reflection coefficients $\Gamma_f = a_3/b_3$ can be defined where b_3 is the power received at the port and a_3 is the power reflected back due to mismatch in termination. Similarly, the reflection coefficient due to the switch as seen in port 1 can be measured with $\Gamma_r = a_0/b_0$. These two reflection coefficients are called switching terms. Note that since the signal flow path in the forward and reverse direction is different these two terms will have different values.

Since the through structure is a passive element it should be reciprocal in nature, i.e., $S_{12} = S_{21}$. But due to reflection from the switch termination, actual measurement will not show this behaviour. Detailed description of switch term correction can be found in [28].

The effect of switch terms can be included in the port match and transmission tracking error terms (see figure 4.7) for both forward and reverse stimulation. In figure 4.7, if the cross-talk terms e_{30} and e'_{03} are ignored, only 5-terms will remain in each direction of stimulation. This is referred to as 10-term error model and can be derived from the error matrices E_A and E_B which were determined using SOLR calibration. Once the 10-terms are known equations 4.7 to 4.10 can be used to calculate actual device S-parameters from the measured S-parameters.

4.3. De-embedding of pad parasitics

Usually the calibration reference plane is defined at the top metal level where the probes connect to the pad or some intermediate layer in the stack. Thus calibration will correct the reference plane for s-parameter measurement at the pad level or some intermediate level. From the perspective of device modelling it is necessary to bring the reference plane to the M1 layer (see figure 4.10) i.e., close to the MOS transistor terminals. The MOS transistor has an extrinsic set of capacitance and resistance which is bias independent. The pad and via structure from the top metal layer to the lowest metal layer introduces additional parasitic elements. If the device is modelled as such the effect of such parasitic will lead to error in the parameter values. Also such a model will be not much useful for the designer. In practice, the MOS transistor model used to build the circuit schematic will contain only the M1 metal layer as this is the minimum required to define a functional device. The final layout of the circuit shall require use of additional metal layers to route the required wires. Such routing invariably will introduce parasitic capacitance. Hence the designer will usually make considerations for such additional parasitic behaviour to optimize the performance of the circuit. Hence, before extraction of model, it is necessary to perform a de-embedding step such that the reference plane of the S-parameters is moved to the M1 metal layer.

4.3.1. Two Step de-embedding

Figure 4.10 illustrates a two step de-embedding method where the reference plane of calibration is defined at the top metal layer. The device sits at the bottom of the stack and is accessed using the via. Usually the transistors used will have multiple fingers. Accordingly two dummy structures are created on the characterisation chip. The gate and drain are the two ports for the device with source and bulk tried together to the ground. This is shown in 4.9.

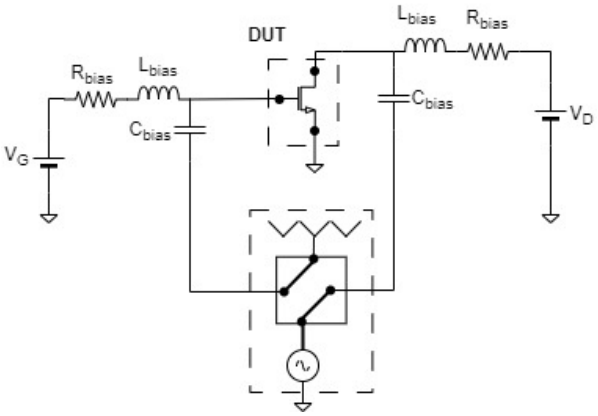


Figure 4.9: Schematic of the basic measurement setup

For an open dummy structure the device is removed while the metal routing remains. To create the short dummy structure the gate and drain terminal metal connections are extended to create a short between the two.

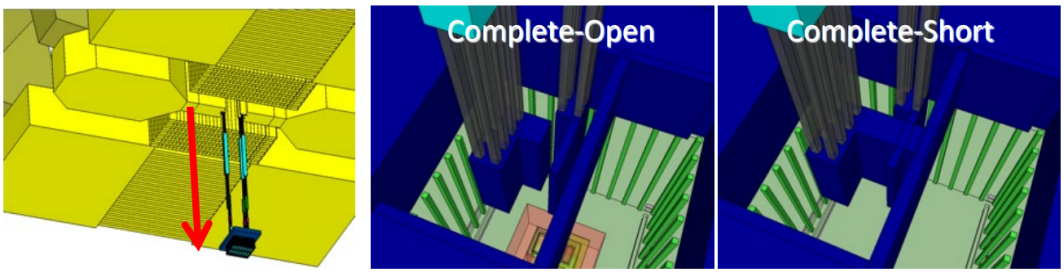


Figure 4.10: Open-Short de-embedding structures [32]

A lumped element approximation can be made to build equivalent circuit as shown in figure 4.11[19]. The admittance Y_{p1} and Y_{p2} are parasitic capacitance to ground and Y_{p3} is the parasitic coupling capacitance between port 1 and 2. In first step the open measurement is used to remove these admittance. In second step the short structure measurement is used to remove the impedance in series with the three device terminals. The calculation steps are shown below. The measured S-parameters of the open and short dummies are converted to Y-parameters to begin with.

1. De-embedd the calibrated Y-parameters of short dummy(Y_{short}) and device structures (Y_{DUT}) with the Y-parameter of the open dummy. Y_{DUT} includes the parasitic admittance and series impedances. $Y_{osc} = Y_{short} - Y_{open}$ and $Y_{doc} = Y_{DUT} - Y_{open}$. This step will remove the parasitic admittance.
2. To remove the three series impedance convert the two Y matrices into Z-matrices and subtract, i.e, $Z_{DUT,actual} = Z_{doc} - Z_{osc}$.

$Z_{DUT,actual}$ is the Z-parameter representation of the device response with reference plane moved to the terminals of the MOSFET. This is shown with black dashed line in figure 4.11. Note that the ground reference is also moved to the device ground. Z_{L3} is de-embeded. This is not shown in the figure.

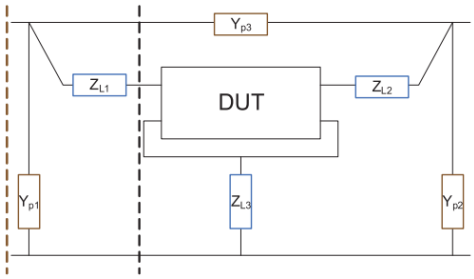


Figure 4.11: Lumped element approximation based equivalent circuit for open-short de-embedding. (image from [19]) The de-embedding step will shift the reference terminal from red line to black line. Figure shows for port 1 but the same is applicable for port 2.

4.3.2. Other De-embedding methods

While open-short de-embedding is used for this work as it is sufficiently accurate up to 40GHz, many other de-embedding methods have been proposed and used over the years. When lumped element approximation is made, a trade off needs to be made on the number of elements used to describe the characterisation fixture and the number of dummy structures required to determine the element values. Open, short and thru dummies were used in [15] and [36] which makes it a three step procedure to shift the reference plane to device terminals. In [27], a four step de-embedding procedure is described using three dummy structures, open short and a combined simple-open-simple-short structure. A multi-line de-embedding technique was presented in [33], [8] where no lumped element approximation was made. These were out of the scope of this work; interested readers can refer to these.

4.4. Reference gamma files from EM simulations

As discussed earlier in order to perform calibration, knowledge of a short, open and load is required. A rough knowledge of the thru standard is necessary in order to calculate all the error terms on the 10-term error model which is used in case of SOLR calibration. These can be obtained from electromagnetic simulations. Hereafter we have shown the structures as on TSMC 40nm technology. Figure 4.12 show the three-dimensional view of CPW transmission line structure formed by the top metal layer (M7 in this case). The device sits at the bottom of the stack and is connected to the top metal layer. The reference plane after calibration will be placed at the inner ends of the transmission lines.

Figure 4.12: Characterisation Fixture without the GSG pads.Only the TL is shown; Pads are not shown here as ADS 3D view ran out of memory to process the structure with pads.

The simulated S-parameter response of the TL is shown in figure 4.13. Ports are placed at the start and beginning of the line on each side. Thus, it will behave like a thru structure.

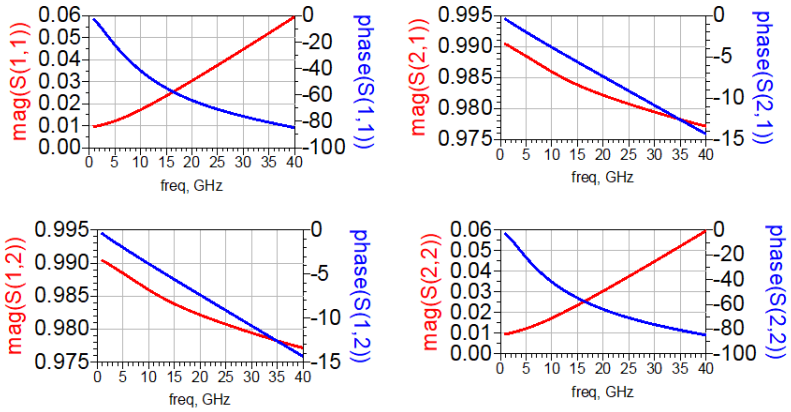
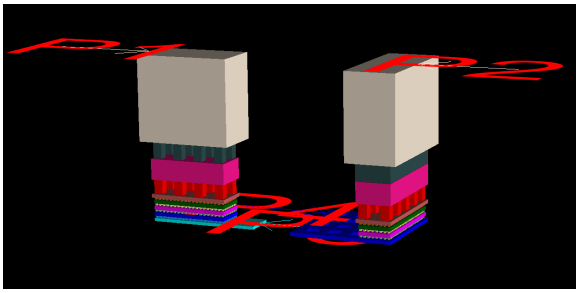
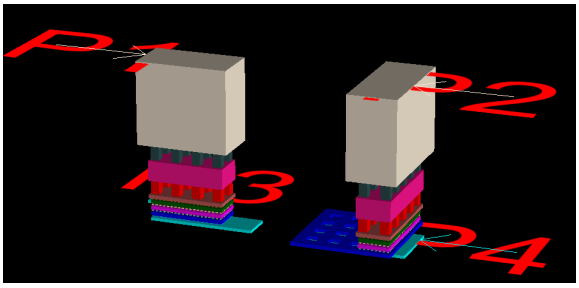


Figure 4.13: S-parameter response of the TL in figure 4.12

4.4.1. Open and Short standard fixtures



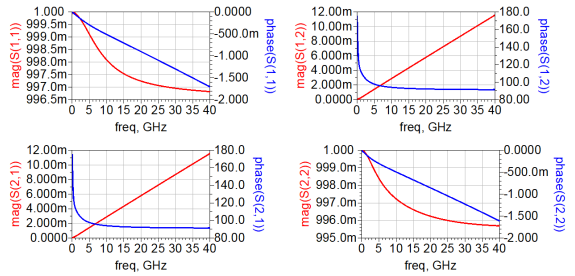
(a) The Open Standard Structure



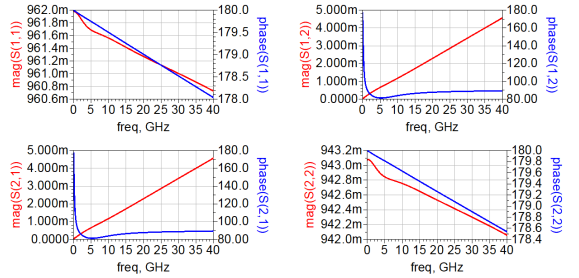
(b) The Short Standard Structure

Figure 4.14: 3D view of the open and short standards. These to standards are very similar in structure. Only difference is that in case of short. M1 layer is extended such that it connects to the ground net. This is not shown here though.

The open and short standard structures are shown in figure 4.14. The reference plane for these are set at the top metal layer M7. To reach M1 layer open-short de-embedding will be used. In case of the open standard the signal path is not connected to the ground net while in case of short the lowest metal layer M1 is extended to the ground net. For EM simulation four ports are defined; P1 and P2 are for input ports. P3 and P4 are the ground pins. In case of open structure P3 and P4 are not connected to ground metal. DC measurement will show open circuit condition. In case of short P3 and P4 are connected to the ground metal. DC measurement will give a very low resistance. The S-parameters obtained from EM model generated in ADS is shown in figure 4.15.



(a) Open Standard



(b) Short Standard

Figure 4.15: S-parameters of the Open and Short Standards from EM simulation model

Note that the magnitude of S_{21} and S_{12} is very small as should be the case as these standards will be used for 1-port calibration. While S_{11} and S_{22} magnitude is about unity meaning reflection of all incident

power. But the phase of S_{11} or S_{22} for open structure is $180°$ out of phase with that of short structures. An ideal short will reflect the incident wave but also cause a $180°$ phase change. The S-parameter can be seen changing with frequency due to the non-ideal nature of on-wafer structures.

4.4.2. Thru standard fixtures

Figure 4.16 shows the thru structure of implementation. The M1 layer is extended from port 1 to port 2. The corresponding s-parameter response from EM simulation is shown in figure 4.17.

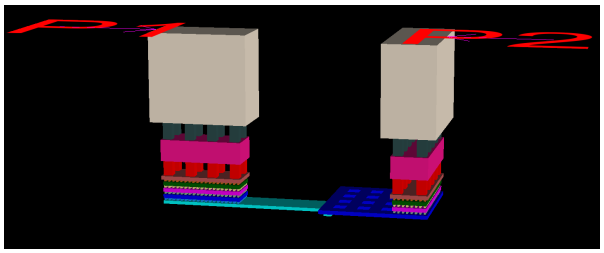


Figure 4.16: Thru standard structure

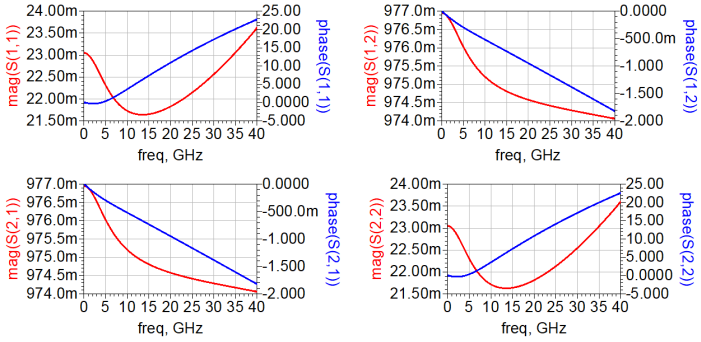


Figure 4.17: S-parameter of Thru standard from EM simulation model

In case of thru standard the S_{12} and S_{21} should be same as it is a passive reciprocal structure. The S-parameter response of the two port thru structure is shown in figure 4.17.

4.4.3. Load standard fixtures

For load standard, each port has a 50Ω n-type poly resistor connected between the port and the ground. For an ideal 50Ω load the reflection coefficient should be zero. The simulated S-parameters of the structure is as shown in figure 4.18.

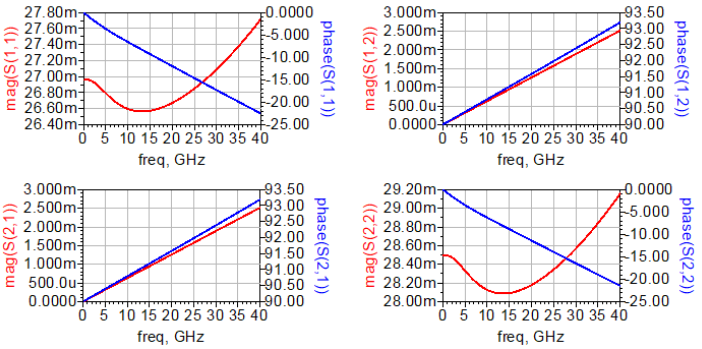
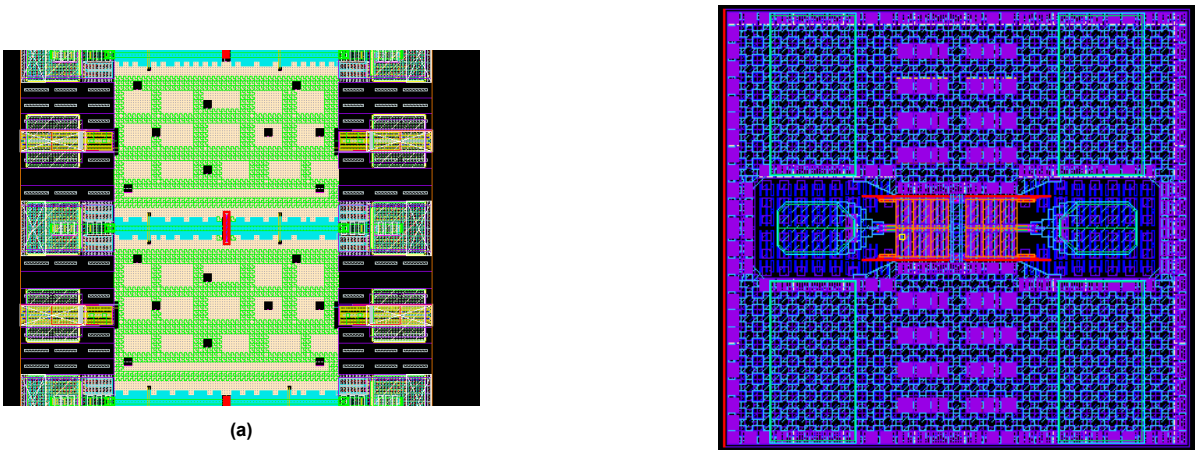


Figure 4.18: S-parameter of the Load standard

For simulation of the load structure, parasitic extraction of the ”rnpolywo” resistor structure was car-

ried out in caliber PEX. Using the generated calibre view s-parameters were simulated. ADS layout tool was not able to map the poly resistor to the correct material definition. As such EM simulated model showed a very low resistance value. Alternatively, only metal connections from the poly resistor to the top metal and ground metal layers were simulated in the ADS EM simulation tool. The s-parameter of the Calibre PEX extracted resistor model was saved in a touchstone file format. Using the S2P component in ADS, the touchstone file was combined with the EM model of the metal connections. This entire structure was then simulated for the final s-parameter of the load standard. This alternative approach may not be as accurate as a complete EM simulation although the reflection coefficient obtained from the used approach is not far from expectation. It is expected that the metal interconnects will introduce additional parasitic capacitance and resistance. The deviation from the ideal 50Ω case can be observed in the simulated S-parameters. In a fabricated chip, it is possible that the load value might show deviation due to manufacturing tolerance. Figure 4.19 shows the layout view of the characterization structure in TSMC 40nm and Global Foundries 22nm technologies. Ground-Signal-Ground pad structure is used as an interface between the probe tip and the characterization structure.



(a) (b)

Figure 4.19: Layout view of characterisation fixture used for RF characterisation

4.5. Measurement Setup

The measurement setup used to characterize the device at various bias points is shown in figure 4.20. The VNA port 1 is connected to the gate terminal side of the characterization fixture and port 2 is connected to drain side. Rhode & Schwarz ZNB40 2 port VNA with in-build bias-tees are used to provide the DC bias from the Keithley 2636B SMUs to the device. For consistency and ease of use SMU-A is connected with bias-tee of port 1 and SMU-B to the bias-tee of port 2.

The SMUs for DC bias voltage source are programmed to provide voltage from 0 to 600 mV for the gate side and 0 to 800 mV and configured for 2-wire measurement. For each bias point the VNA acquires the scattering parameter response from 100 MHz to 40 GHz at 201 frequency points.

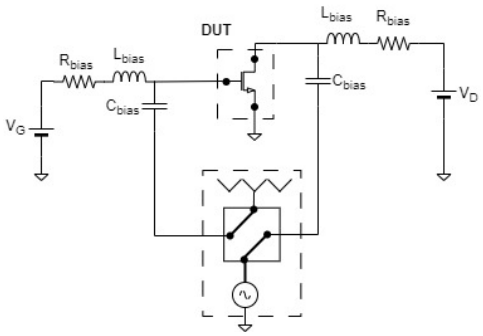


Figure 4.20: Equivalent circuit for the experiment Setup

Since the devices are embedded in a transmission line structure and the bias tee, cables, and cable-port junctions add resistance to the path, it is necessary to determine this external resistance at each terminal of the device. This is necessary in order to determine the DC bias voltage at the device terminals.

This can be done by measuring the short dummy structure accompanying each device. The main contribution will be from bias-tee. Note that it is difficult to get repeatable values of contact resistance on each landing of the RF probes. In such cases, multiple landing and measurements can be done to reach the most likely value. But it should be noted that the pad, which is made of a soft metal like Aluminium, can be worn out and the contact resistance may increase. The contact resistance ambiguity will be mostly observed in the case of landing on open and transistor characterization structures as the resistance cannot be measured in these cases. For short and load structures, a DC measurement can be performed to determine contact resistance.

4.5.1. Considerations for measurement at cryogenic temperature

Since the device temperature will increase during measurement, the bias current is measured before and after the scattering parameters are acquired from the VNA. Measurement time is defined from the moment DC bias is applied till DC measurement after S-parameter measurement. This includes the time required for the SMUs to measure the dc voltage and current. The measurement time for each bias point is about 30 seconds for 201 point frequency sweep.

The samples are mounted on a copper sample holder with GE varnish-based heat-sink glue which is known to be usable at cryogenic temperature. The effect of self-heating the device during measurement is possible. The nearest temperature sensor is underneath the sample stage of the probe station and needs to be kept at a steady temperature. Also, a Helium dewar of 100L can provide limited time for measurement which was observed to be about 6 hours at steady 4.2K sample stage temperature. In order to keep the temperature constant at 4.2K the dip-stick needs to be lowered periodically in this setup to maintain helium flow.

The DC and RF measurements are taken with as close time proximity as possible for device model consistency. RF response of the device is dependent on the bias condition and temperature of the device. A large time difference between DC bias current measurement and S-parameter measurement may lead to error as the device temperature may change. For cryogenic temperature measurement, this needs to be considered as the device may heat up to a higher temperature rendering the measured data point inconsistent.

5

Measurement Results and Discussion

Based on the procedure developed in the previous chapters, CMOS transistors of various dimensions from Global Foundries 22nm FDSOI technology were measured at room temperature and cryogenic temperature. The measurement results and the extracted parameter values are shown hereafter.

5.1. DC Measurement Results

Technology	**GF 22 nm FDSOI**	
Nominal $\lvert V_{DS} \rvert$	800mV	
Nominal $\lvert V_{GS} \rvert$	600mV	
W/L (nm/nm)	$500/18$ $1000/18$ $300/40$	$500/18$ $1000/18$ –

Table 5.1: Summary of device size measured

The devices mentioned in the table 5.1 were measured in room temperature and at 4K using Lakeshore make cryogenic probe station. The setup is shown in figure 5.1.

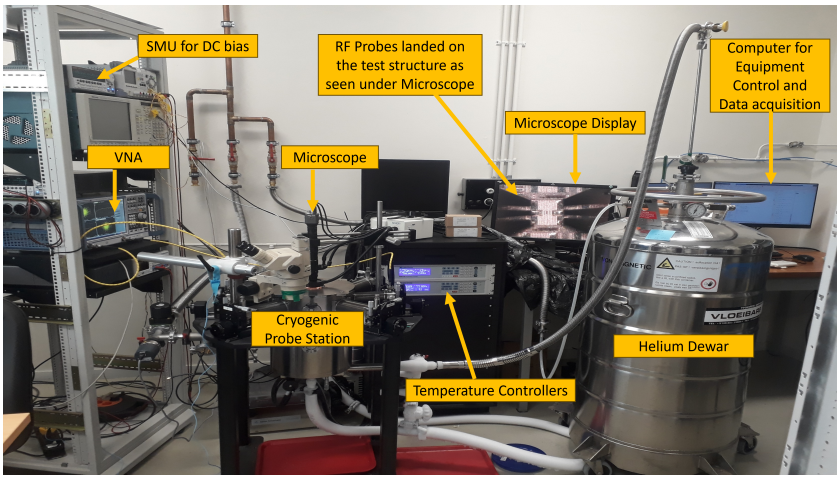
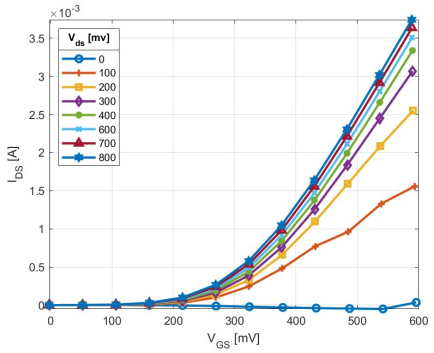
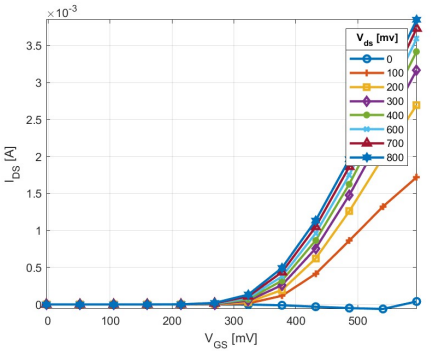


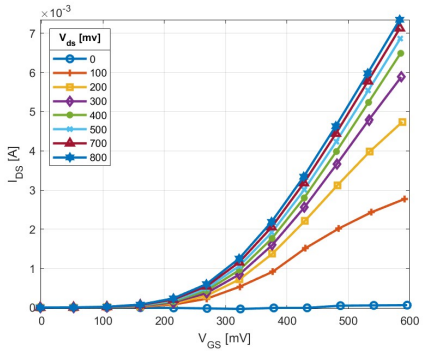
Figure 5.1: Equivalent circuit for the experiment Setup



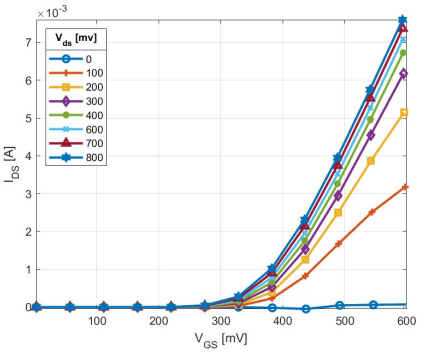
(a) NMOS W = 500nm; L = 18nm, nf = 16 at 295K



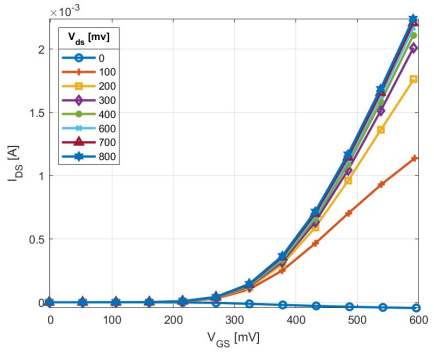
(b) NMOS W = 500nm; L = 18nm, nf = 16 at 4K



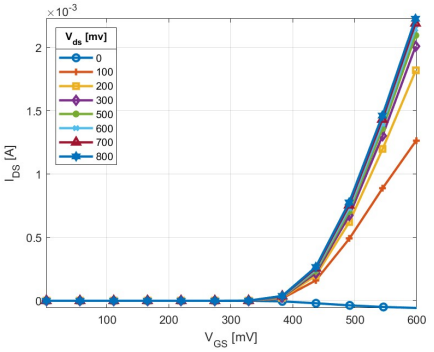
(c) NMOS W = 1000nm; L = 18nm, nf = 16 at 295K



(d) NMOS W = 1000nm; L = 18nm, nf = 16 at 4K

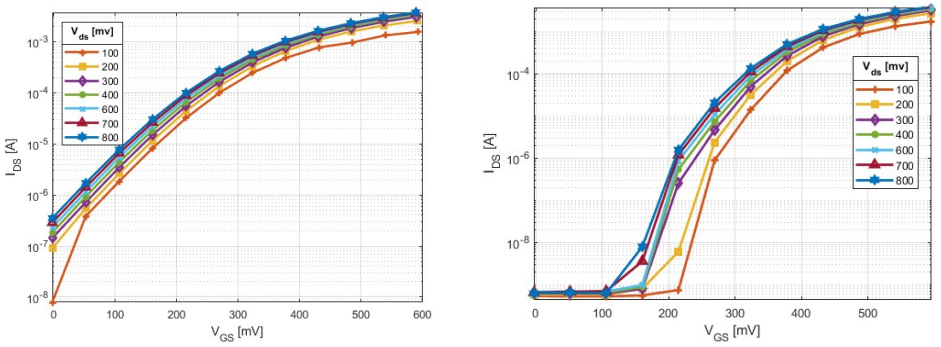


(e) NMOS W = 300nm; L = 40nm, nf = 8 at 295K



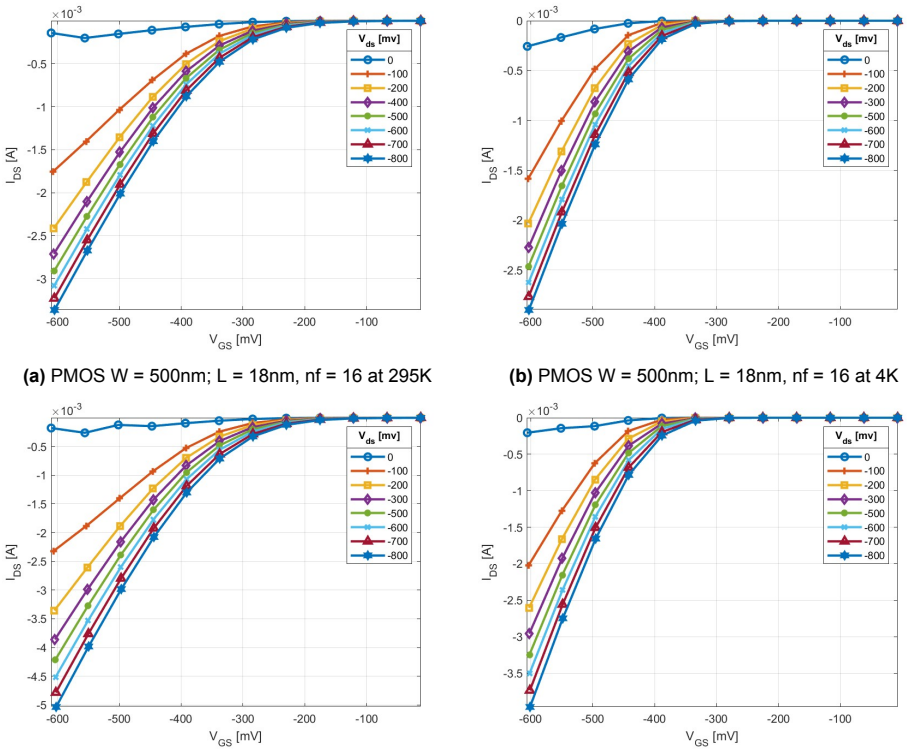
(f) NMOS W = 300nm; L = 40nm, nf = 8 at 4K

Figure 5.2: DC Measurement: I_{ds} vs V_{gs} Plot for NMOS transistors at room temperature and 4K. V_{DS} shown here are rounded up to show the DC IV trend. $V_{DS} = 0mV$ is actually less than 0 mV. These measurements were captured in parallel with S-parameter measurements. For each of the curves, V_{DS} shows very small variation possibly due to two-wire measurement. To show the shift in threshold voltage and the general trend of the current, rounding-up is done.



(a) NMOS W = 500nm; L = 18nm, nf = 16 at 295K **(b)** NMOS W = 500nm; L = 18nm, nf = 16 at 4K

Figure 5.3: DC Measurement: I_{ds} vs V_{gs} Plot for NMOS transistor at room temperature and 4K. With y-axis on log scale the effect of temperature dependence on threshold voltage is visible.



(a) PMOS W = 500nm; L = 18nm, nf = 16 at 295K **(b)** PMOS W = 500nm; L = 18nm, nf = 16 at 4K

(c) PMOS W = 1000nm; L = 18nm, nf = 16 at 295K **(d)** PMOS W = 1000nm; L = 18nm, nf = 16 at 4K

Figure 5.4: DC Measurement: I_{ds} vs V_{gs} Plot for PMOS transistors at room temperature and 4K.

From figure 5.2 and 5.4, the threshold voltage shift can be observed at 4K. Figure 5.3 shows the 500nm/18nm NMOS transistor DC behaviour with current axis in log scale. The effect of temperature on the threshold voltage is more visible. For the case of NMOS devices, an increase in current is observed. But in the case of PMOS, while threshold voltage shift is visible, drain current has rather lower values. But note that the threshold voltage has increased considerably too. Since the measurement was carried out in the nominal voltage range up to 600mV, the crossover point was not captured. In figure 2.5 (a), it can be observed that as the gate voltage is extended till 1 V, the drain current at low temperature is higher than that at room temperature. The DC measurement of NMOS shows reverse current flowing from source to drain in case of low V_{DS}. It was observed from DC measurements on ISS short structure, that the VNA port ground was not at 0V. This is shown in figure 5.5. Short Dummy structures on the wafer can also be used.

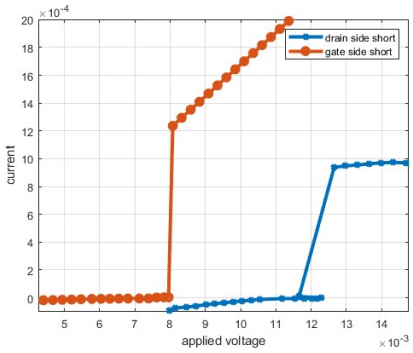


Figure 5.5: DC offset measurement on ISS Short structure. Current is in ampere and voltage in volt

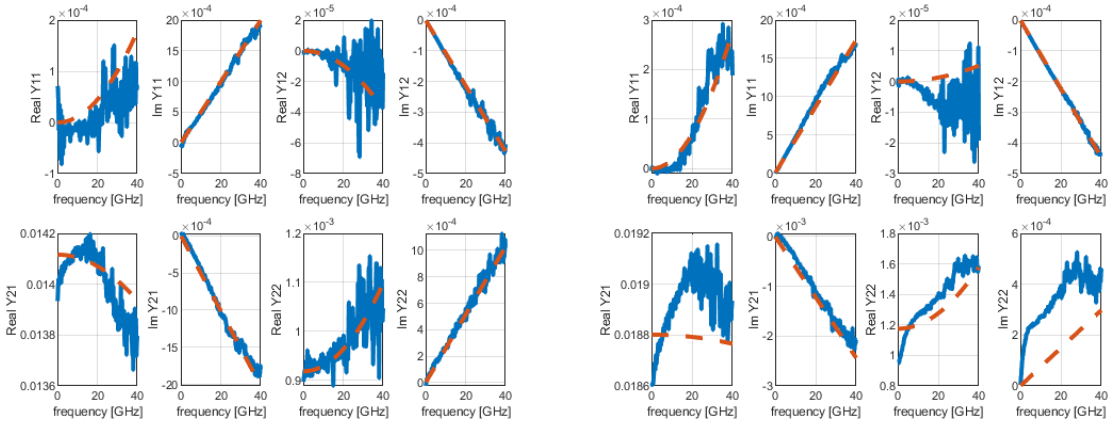
When drain voltage bias lower than this offset is applied through the probes the actual voltage seen at the device terminals is negative. In addition, access resistance will also add voltage drop. Thus, actual V_{GS} and V_{DS} will be different. Also, since two wire measurement was carried out the measured DC voltage varies by a very small figure usually less than 1mV. In the IV plots shown here, the drain voltage in the legend is rounded up to enable I vs V plotting. The plots look noisy as such. Also, these voltages are corrected for path resistance to get the actual terminal voltage. The $V_{DS} = 0$ is indicative only; actual values are small negative numbers due to the mentioned offset and hence the negative currents in the plot.

5.2. RF Measurement result

The DC measurements were performed in parallel with S-parameters measurements. Thus, even if the device channel is at a considerably higher temperature than the ambient, both measurements are consistent with each other with respect to temperature.

Figure 5.6, shows the measured Y-parameters obtained after calibration and de-embedding of raw measurement data at $V_{gs} = 589mV$ and $V_{ds} = 753mV$ for an NMOS transistor. These corrected voltage figures are different from what is applied by the SMU. In order to do these DC voltage drop corrections, one can use the short and open dummy measurement data after calibration.

From the first step of de-embedding (see de-embedding discussion in the previous chapter) the open corrected short structure Y-parameter response is available. Converted to Z-parameters, the real part will have path resistance till the M1 layer. Extrapolation of $Z_{11} - Z_{12}$ to low frequency gives the resistance from probe tip to gate side M1 layer. Similarly, we can get drain side access path resistance from $Z_{22} - Z_{21}$. Z_{12} and Z_{21} will give the source terminal resistance. The DC resistance from SMU to probe tip can be obtained by DC IV measurement on ISS short structure or gold landing pads. Note that contact resistance will introduce some ambiguity. But a few iterations of short dummy measurements will give an indication of the most likely resistance values. Note that these values can also be determined by measuring the short dummy structure but in that case, contact resistance variation can lead to ambiguity.



(a) Measured at 295K

(b) Measured at 4K

Figure 5.6: Measured Y-parameter for NMOS W = 500nm, L = 18nm, nf = 16 at $V_{gs} = 589mV$ and $V_{ds} = 753mV$

The measurement showed noise, especially in the real part of the Y-parameters. Since 201 points in frequency were captured for the S-parameter measurements, the Y-parameter shows a clear trend that can be used to extract model parameters. This can be done by determining the best-fitting curve. In this work, the ”fitlm” function in Matlab was used. This is similar to the IRLS method used in [23]. But note that any such function will be limited by the noise in the data, which in this case was excessive.

The root cause of this noisy response was not understood. Interference from any other instrument or from a power supply can be ruled out as the IF bandwidth was set to 50Hz. This behavior was consistently observed over multiple measurements. The VNA ports were tested with available Anritsu cal-kit in the 10MHz to 40GHz range which didn't show any anomalous behaviour. Also, multiple measurements were made on ISS CS-5 cal-kit but distorted behavior was found in all cases. Thus, intuitively investigation of probe arm assembly is the logical choice. Averaging of the S-parameter measurement did not give improvement. Due to time limitations, further investigation of the source of distortion was not carried out in this work. The modeling approach using ANN can still be investigated by determining the best-fitting curve. Although the accuracy of the model cannot be determined as the data itself is distorted. Figure 5.7 shows the best-fit curve determination. The black dashed lines are the initial guess that is used by the ”film” function. The red line is the best-fitting curve determined by the function. The imaginary part of Y-parameters is ideally straight lines with constant slopes. Thus, the function will fit this straight line overlapping a maximum number of data points. The further away the point is, the lower the weight it gets in the fitting process. The same is applicable with real parts but in this case, the fitting curve should have quadratic dependency on frequency. The initial approximation has a quadratic dependency on frequency and the curve fitting process will maintain this while readjusting the coefficient of the quadratic term to get the best fit.

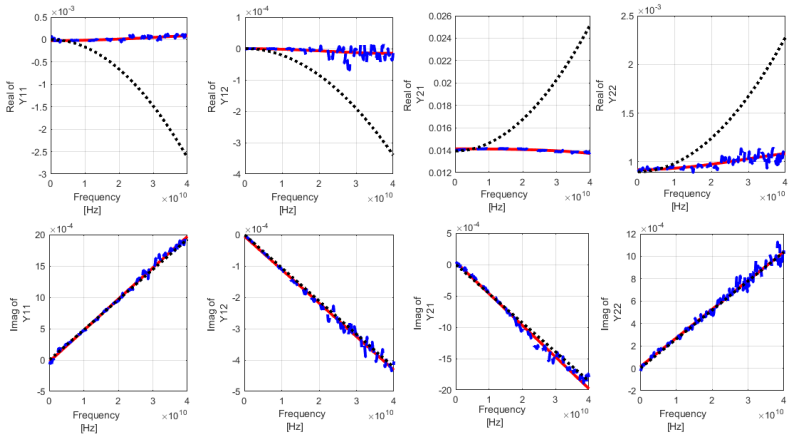
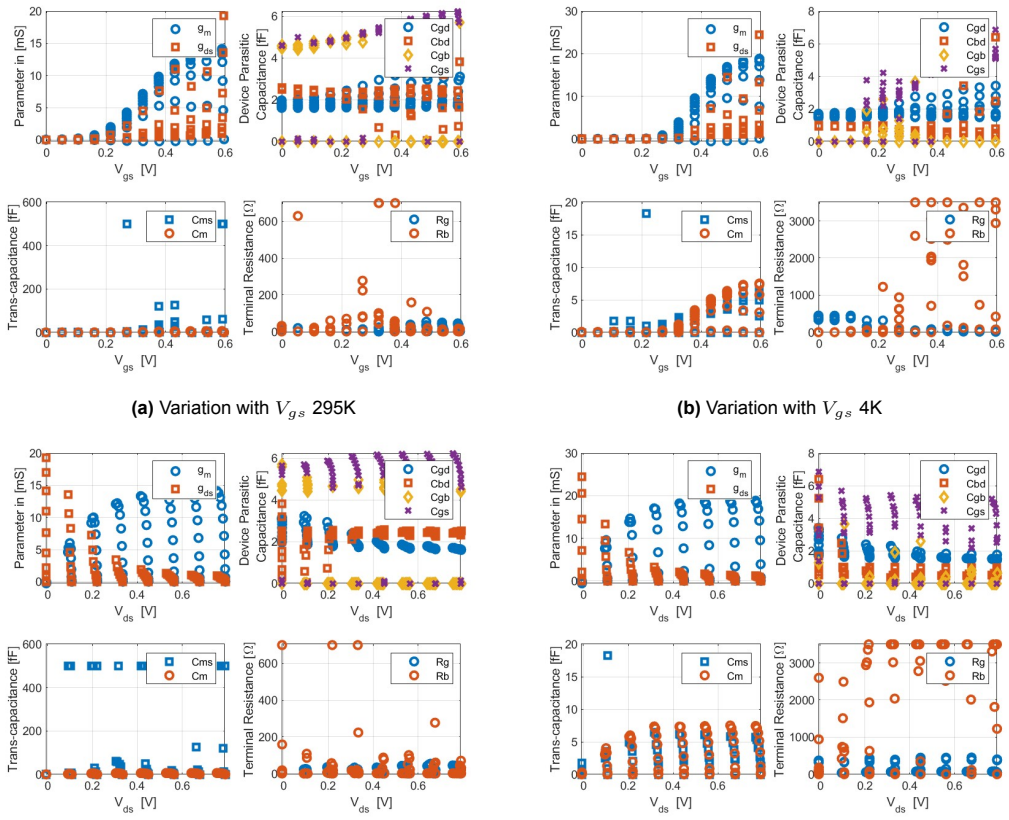


Figure 5.7: Determination of best-fitting curve from measured data. In this plot RT data is used. Red solid line: Best fit curve. Blue dash: Measured data. Black dashed line: First approximation to trigger curve fitting. The measured real part is noisy but due to the plot scaling, it is not clearly visible. Transistor size is same as in figure 5.6

5.3. Extracted Parameters



(a) Variation with V_{gs} 295K

(b) Variation with V_{gs} 4K

(c) Variation with V_{ds} 295K

(d) Variation with V_{ds} 4K

Figure 5.8: Extracted parameters for NMOS W = 500nm; L = 18nm, nf = 16.

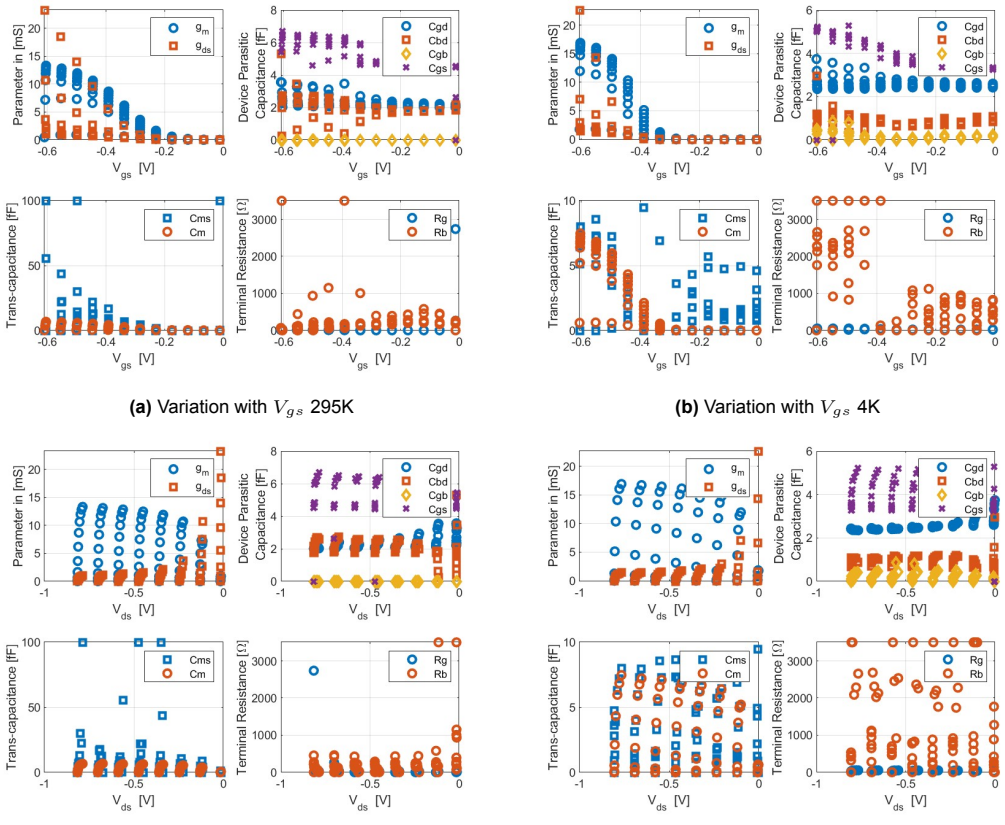
The variation of the parameters with terminal voltages is shown in figure 5.8 for a NMOS transistor (W = 500nm; L = 18nm, nf = 16). The extraction is done from the Y-parameters of the transistor after open-short de-embedding. These parameters are extracted analytically using the method described in section 3.4.1. The values of g_m, g_{ds}, the capacitors and resistors in the model are similar to what

we got from room temperature simulations in Chapter 3. Other than the outliers in the extracted data, which is due to the noisy data, the values are in the same range.

Threshold voltage shift is noticeable from g_m and g_{ds} vs V_{gs} plots (see (a) and (b)); see how the values are near zero below 200mV in 4K. In strong inversion and saturation region, a significant increase in g_m is observed at 4K ($\sim$ 20 mS) due to increased mobility, compared to Room Temperature($\sim$ 15 mS). C_m and C_{ms} do not show much change between the two temperature.

R_g and R_b show higher values at 4K. But observing from plots (b) and (d), in strong-inversion and saturation bias, the resistance values are actually approaching very low values as expected. In [23], the R_b and R_g were reported to be very low at cryogenic temperature. R_g is the resistance of gate polysilicon. Due to reduced scattering at 4K the resistance values should be lower. The reason for this is that the frequency-dependent coefficients of the fitted Y-parameter curves are such that it forces R_g to take large values. The fitting operation is done independently on each of the curves. But the analytical solution is based on a first-order approximation of the device Y-parameters which are correlated to each other through the equivalent model parameters. Due to the noise in the Y-parameters, which we observed to be excessive in some cases, the best-fit calculated is also not preserving the correlation. This is not a limitation of the curve fitting function but rather we are limited by the noise in data. While describing the analytical extraction procedure this was something we anticipated.

The value of C_{gs}, C_{gd} and C_{gb} show minor changes between 4K and RT for gate voltage above threshold voltage. At 4K, C_{bd} shows some drop in its value at low Vgs compared to RT. In strong inversion, the value is similar to RT. In general, the capacitance values do not show much of sensitivity to temperature. Both [12] and [23] have reported this behavior. This is also apparent from the imaginary part of the Y-parameters shown in figure 5.6.



(a) Variation with V_{gs} 295K

(b) Variation with V_{gs} 4K

(c) Variation with V_{ds} 295K

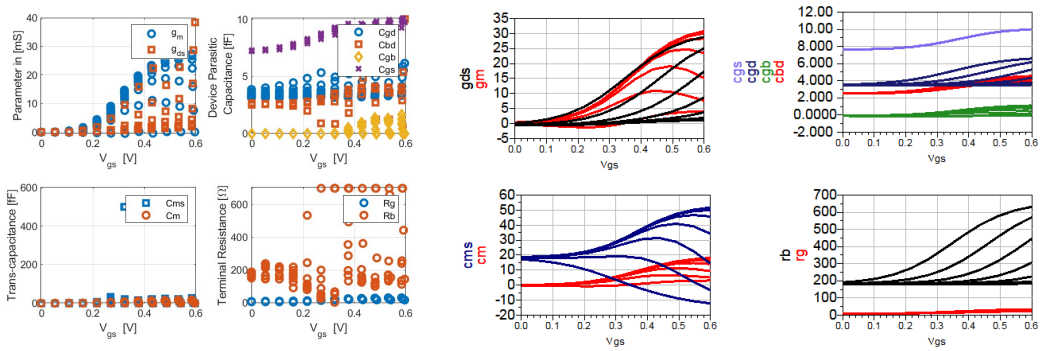
(d) Variation with V_{ds} 4K

Figure 5.9: Extracted parameters for PMOS W = 500nm, L = 18nm, nf = 16.

Figure 5.9 shows the extracted parameter variation with terminal voltages for a PMOS device (W = 500nm, L = 18nm, nf = 16). The capacitance values do not show much deviation except C_{bd}. R_b shows some increase at 4K as in the case of NMOS before.

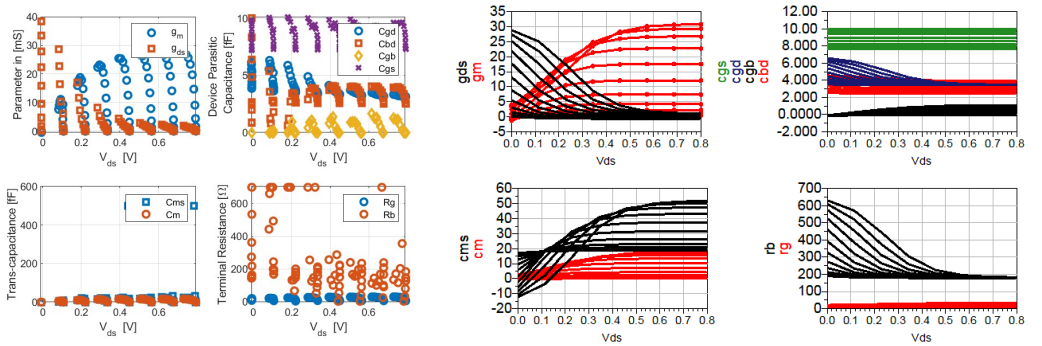
5.4. ANN Model Generation

The extracted parameters were used as training input to build ANN model of the transistors using ICCAP. The training configuration was set for six inputs (V_{gs}, V_{ds}, W, L and T). The number of hidden layers was set to two and for each hidden layer the number of neurons was set to six. The equivalent circuit was implemented in Verilog-AMS. SP simulation of the Verilog implementation was performed at various bias points. The ANN generated function calculates the parameters for each bias point. The verilog-AMS code uses these calculated parameters for SP simulation. Figure 5.10 shows the parameter values extracted and the parameter calculated by the ANN generated function for 1000nm/18nm NMOS transistor for room temperature.



(a) Extracted Parameter vs V_{gs}

(b) Extracted Parameter vs V_{gs}



(c) Extracted Parameter vs V_{ds}

(d) Extracted Parameter vs V_{ds}

Figure 5.10: Extracted parameters and those calculated by the ANN generated model function for NMOS W = 1000nm, L = 18nm, nf = 16 across bias range

g_m was modelled in close agreement. At $V_{gs} = 584mV$ and $V_{ds} = 728mV$, g_m from ANN model is 30.4 mS while the measured value is 27.22mS, thus an error of 11.68%. But for g_{ds} the values are 1.108 mS and 1.908mS respectively, thus 41% error. Similar observations can be said for the terminal capacitance. But note that in the extracted parameters the trans-capacitance values had some outliers. The ANN training did not include those. This is what we wanted to do in the first place and shows the advantage of using this approach. C_m is well captured in the model as well. An outlier for C_{ms} is observed around 400mV of Vgs in plot(a). These outliers were rejected and a more gradual variation with voltage bias is observed in the values calculated by the ANN-generated function. The same can be said for R_b and R_g. Note that C_{ms} from the trained model shows some negative values too, especially at low-bias voltages. The reason for this is two-sided. The given dataset is limited in number of data points at low bias conditions. Parameter variations are more at such low bias conditions. Thus, resolving this issue will require adding more data points in weak inversion region, then resizing the ANN and the number of training iterations to find the optimum model function. The ANN model is in general coming in close proximity to measured values though. Due to limited time available, number of iterations were made to size the ANN were limited.

The Y-parameter simulation of the Verilog implementation of the small signal model of 500nm/18nm NMOS transistor is shown in figure 5.11. The ANN was configured with 2 hidden layers with 30 neurons in each layer in this case. This size is different from used before and was obtained after multiple trials.

The maximum iteration for training was set to 2000. The model generate with this size 2000 iterations showed a close fit to the measured data.

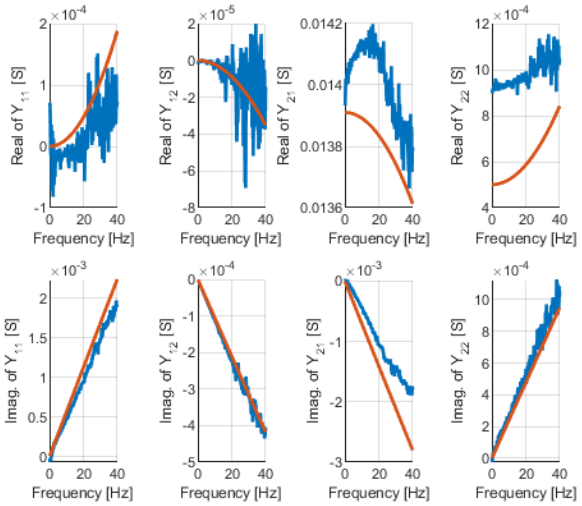


Figure 5.11: Y-parameter of NMOS (W = 500nm, L = 18nm, nf = 16) under strong inversion and saturation bias($V_{GS} = 589mV$, $V_{GS} = 753mV$). The blue line is measured in RT and the orange line is Verilog-AMS implementation. SP simulation was done in ADS.

Variation was also observed between the two training cycles. This is due to the fact that the initial weights and biases in the ANN are randomly initiated. Based on training, these are adjusted. If the number of training iterations are kept small one needs to rerun the training cycle a few times to obtain a better ANN model. In each run, the weights and bias so calculated will be different from the previous run. A large training iteration can lead to overfitting of the dataset, thus the outliers will also be included in the ANN model.

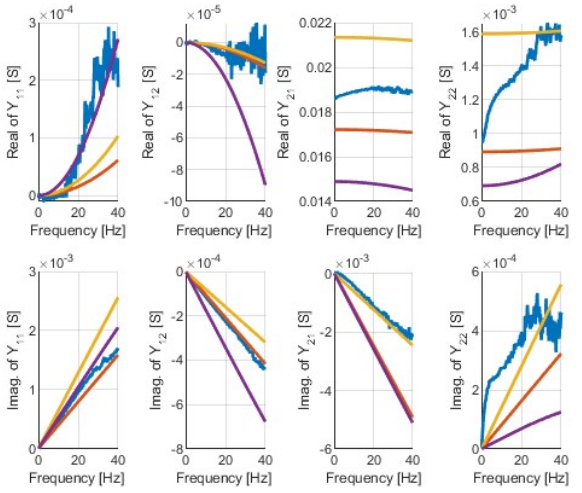


Figure 5.12: Y-parameter of NMOS (W = 500nm, L = 18nm, nf = 16) under strong inversion and saturation bias($V_{GS} = 595mV$, $V_{GS} = 767mV$). The blue line is measured in 4K and the orange line is Verilog-AMS implementation. SP simulation was done in ADS.

Figure 5.12, shows the measured Y-parameters and the ANN model simulation in ADS for 4K. The voltage bias is close to figure 5.11. The purple, yellow and red lines are simulations from three different training attempts performed sequentially. In each attempt the number of training iterations was increased while the ANN size was kept as two hidden layers with 30 neurons in each layer, same as

previous plot for room temperature. The values calculated by the ANN model tend to stay in proximity to measured values. Thus few more attempts would have given better results.

5.5. Discussion

5.5.1. Comment of TSMC 40nm Characterisation

Before starting RF characterization, DC functionality check is performed. This helps in the identification of any shorts or no connections or damage in general while gluing the chip to the sample holder. The TSMC characterization chip failed this check. It was found from measurements that a low conductivity path was present between the V_{DD} pad, which is meant to reverse bias the n-well, to the ground pads on the chip. Upon thorough investigation of the layout it was found that at six locations, a metal wire in the M5 layer was misplaced which caused a low resistive connection between the M5 layer of the MOSCAPS (which should stay at high voltage) connects to the M5 layer in the ground pads (which should be at ground voltage). As such the substrate was lifted to a higher voltage than the ground at one end of the chip while at the other end, the ground pad side it was connected to DC ground. This is basically the conductive path between V_{DD} and ground and is undesirable. As a result of this when V_{DS} was low compared to the lifted substrate voltage, a reverse current from the source terminal to the drain flows. The DC IV measurement showed a voltage offset due to this as shown in figure 5.13. We also noticed that at 4K where the substrate will face carrier freeze-out, this offset is not there. As proper DC functionality is required for RF characterization, this chip was not measured any further.

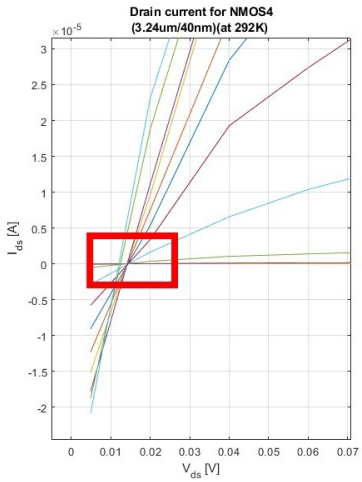


Figure 5.13: Offset in DC I_{DS} vs V_{DS} curves due to current through the substrate of the TSMC chip.

5.5.2. The problem of repeatable contact resistance

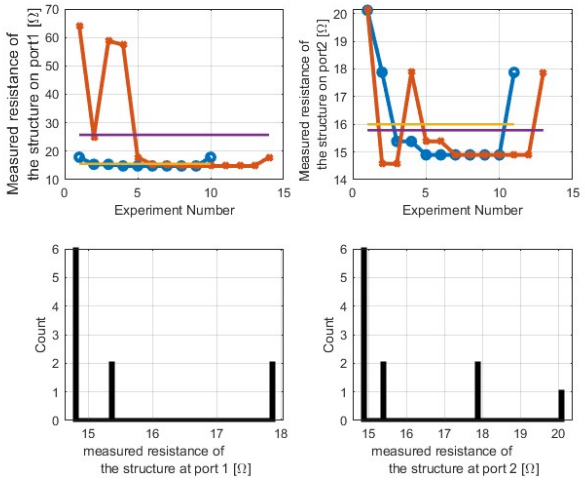


Figure 5.14: Measurement of resistance due to bias-tee and cable using the Thru structure on the chip. Blue and Red lines are two different measurement sessions.

For the manual probe station used in this work, it was observed that contact resistance will vary between two landings. In figure 5.14, the resistance measured on thru structure is shown. This resistance will include the DC bias, cable, and contact resistance between the probe and pad. The blue and red lines show the measured resistance for different measurement sessions. In each session, multiple landing of the probe is performed and the resistance is measured. Also, the power source is changed from port 1 to 2 or otherwise. Note that there are a number of outliers in the measured resistance. This is due to poor landing in such cases. Note that the values start converging to about 15Ω after a certain number of experiments. This can be seen in the histogram plots for each of the ports where six experiments showed resistance very close to each other.

It was observed that the landing gear on each probe arm gets rigid as the probe station is put in a vacuum for cryogenic measurements. Thus the problem of contact resistance will be even more as landing the probe itself is more challenging. In this work, we did similar measurements on short and load structures to determine the most likely values. We use them as a reference while landing the probe at cryogenic temperature. From a practical perspective, we observed that about 15 minutes of wait time after the probe station temperature has reached 4 kelvin is required to make good contact between the probe tip and pad. Possibly this might be due to the thermalization of the probe tips via the thermal anchors.

5.5.3. General Modeling Flow

Based on the analysis carried out in the previous chapters and the results presented in previous sections the modeling flow based on a small signal equivalent circuit using ANN as the tool to generate the model is shown in figure 5.15.

S-parameter and DC voltage-current measurements of the device are corrected for systematic error induced due to non-idealities of the measurement setup. This is where SOLR calibration or as a matter of fact any other calibration algorithm will be used. The DC measurements are corrected for the voltage drop due to path resistances to get actual bias voltage across the device under test. The corrected data is used to extract device model parameters at various bias points. A small signal equivalent circuit, which can replicate the device behavior in the desired frequency range of operation, and a parameter extraction flow corresponding to it is required as a prerequisite. Depending on how detailed the equivalent circuit is, the model will be able to replicate the behavior with higher accuracy.

Y or Z -parameter equations of the equivalent circuit becomes rigorous as the model complexity increases. Thus a trade-off is required between how accurate the model should be and the complexity of parameter extraction. Although it is possible to develop a multi-parameter optimization process these often come with the requirement of providing good start points. Also, upper and lower bounds are

required to keep the values reasonable. From the discussion of the small signal model used in this work, it is known that real and imaginary parts of Y-parameters can be expressed as polynomial functions with frequency as the independent variable. Depending on the order of the polynomial required to fit the data to a reasonable accuracy one can select the complexity of the small signal model.

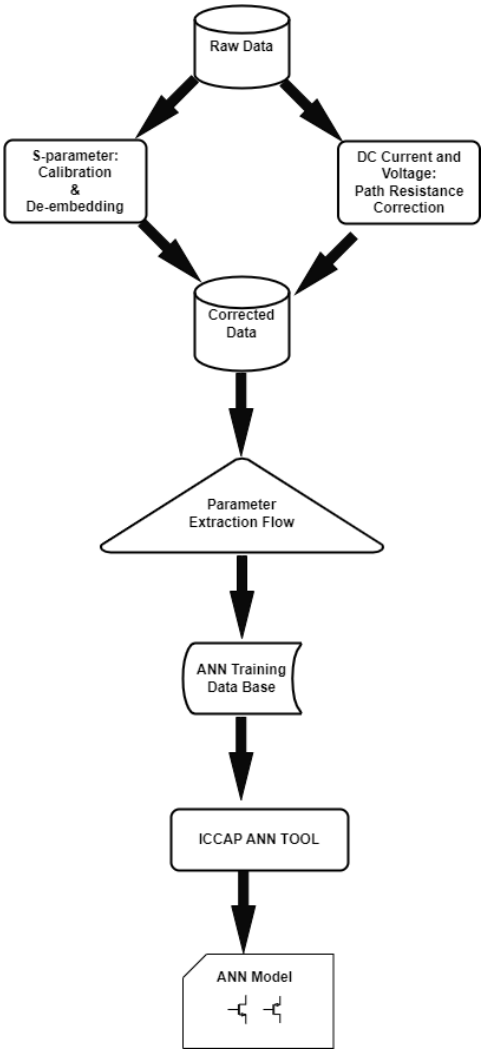


Figure 5.15: General Modeling Flow

5.5.4. Comments on ANN-based approach

The ANN-generated function reduces the, otherwise considerable, effort required to generate equations that relate DC bias voltage and temperature to the parameter values. One demerit of using ANN based approach is that the intuitively sized ANN might not be able to generate a sufficiently accurate model function. Multiple iterations are required to reach the best case size of ANN and the number of training iterations required. This is the reason why there is an offset between the measured Y-parameter and ANN model simulations. We saw in the extracted parameter that the values can lose physical significance to get the best fit with respect to the measured data. This is what we anticipated. Two things need to be mentioned here. First, in the analytical extraction process, no optimization step is available. We do not set any boundary conditions for the parameter values. Second, in the curve fitting step used to get best-fitting curves on the noisy measured data, no optimization of the final Y-parameter values was done. Note that these Y-parameter values are correlated with each other as can be understood from the simplified first-order analytical expression. Curve fitting does not guarantee that this correlation will remain.

Depending on the configuration used for ANN training, the model might overfit or underfit the data set. Iteration with different configurations is required to reach a reasonably good fitting curve. While

this is a challenge when a unidirectional flow for model generation is used, a denser data set can be used to generate the model.

5.6. Conclusions

From the RT and 4K measurements, we observed that in general, the capacitance values show very limited variation between the two temperatures. The major change is shown by transconductance and channel conductivity at 4K. This is due to an increase in mobility. The extracted R_g shows some increase in weak inversion in 4K temperature. The possible reason for this is the curve-fitting on the noisy data. As explained earlier, the coefficients in the real parts of the Y-parameters are correlated but curve fitting will not preserve it as such the values so calculated takes unrealistic numbers.

Even with these limitations, we have successfully shown that the extracted parameters can be used to model small-signal equivalent circuits for CMOS transistors using the ANN modeling approach. We showed how ANN training was able to remove outliers in the dataset.

6 Conclusion

Cryogenic electronics plays a crucial part in realisation of practical Quantum Computer. The design of control and readout circuits, based on CMOS technology, for quantum processors requires the availability of device model valid at extremely low temperatures of 4K or below. While the physical operation of a CMOS transistor has been investigated significantly over the years, an equivalent model that can be used for circuit simulation at 4K is still not widely available. An approach using ANN to close this gap has been explored in this work.

6.1. Summary of findings

A small signal equivalent circuit that can replicate the electrical behavior of the CMOS transistor in 22nm FDSOI was reviewed, to begin with. The equivalent model parameters can be found analytically. This was possible due to simplifications made to the general model. The model was first studied with PDK simulations up to 40 GHz. The extracted parameters from simulation data were then used to generate ANN models. This stage gave us insights into what to expect from this modeling flow. In this process, we also realized that if the extraction algorithm is used for the entire bias range, we will get parameter values that are not necessarily justified from physical understanding.

The previous step showed promising results and thus gave us the confidence to extend it to real measured data. The extraction process was repeated with measured data and the ANN model was generated using that as the input dataset.

The advantage of this method is that no rigorous equations need to be derived as the ANN can do that from the training data set. Using measurements at room temperature and at 4K, it was shown that the ANN-based approach can be used to build a transistor small signal model. We further implemented the small signal model in Verilog-AMS. Thus, allowing us to carry out small signal analysis of circuits.

With the advantages of using ANN as a model generation tool, some major disadvantages were also noticed. The ANN size selection was done by trial and error method. But it is a very tedious and usually non-intuitive process. Too large a size of the ANN can cause over-fitting and is not feasible in practical scenarios where training data will have scattered values.

A general modeling flow was presented that uses the ICCAP ANN tool to generate the model. This model, which is basically a function mapping the input to device parameters, can be used directly in Verilog-AMS.

Note that we were able to use the same set of procedures at cryogenic temperature for model generation. We didn't require any physical equations that relate the behavior to temperature. The only set of knowledge was the expected change which is available from various characterization efforts over the years.

On the experimental side, the SOLR calibration was reviewed. Thermal considerations were taken while preparing the sample for measurements. The measurement results show the expected increase in saturation current and transconductance. In a manual probe station, like the one used in this work, which goes to 4K using an external source of liquid Helium, it will be challenging to keep the temperature constant over a long duration of time. Note that helium dewar will allow you to keep the setup in 4K for a limited time period only.

We also showed that contact resistance repeatability is an issue, especially with open and transistor structures on the chip as we cannot judge the landing in such cases. But in the case of load, short, and thru structures we can use statistical analysis of multiple measurements to reach the most likely DC resistance value at room temperature. At cryo-temperature, we can use these values as a reference when measuring these devices as the resistance will not change drastically.

To summarise, this work establishes the feasibility of using ANN as a model generation tool. Some aspects of this model generation flow need to be developed further to improve the model accuracy and efficacy of the ANN tool. The analytical extraction technique is very useful for frequency ranges up to 40GHz as in this work. Also, model generation can be done for small signal-equivalent circuits with any number of parameters. Thus, the frequency range of validity can be extended as desired.

6.2. Future work

Based on the observations made during this work, a method to effectively select the size of ANN needs to be explored in the immediate future. This will be crucial for model generation as the accuracy of the ANN model will be limited otherwise. In the future, we can increase the number of data points. This will also be beneficial for better ANN training. Temperature sweeps can also be considered.
The small signal model needs to be upgraded further by allowing more components in it. This will require recalculating the analytical expressions of the Y-parameter representation. The use of boundary conditions during the extraction of parameters is required. This will solve the problem of unrealistic parameter values. We noticed that noise was present in the measured S-parameters. The source of the noise and possible methods to eliminate it needs to be investigated.

We didn't measure the device temperature accurately as on wafer temperature sensing element was not available. In the future, this can also be implemented to capture heating effects. The use of ANN to model that will be an interesting possibility to explore. Measurement and data acquisition can be linked to the measured temperature of the device using the sensing element and the measurement acquisition can be automated to allow intermediate cooldown of the sample. As for the experimental setup, methods to improve contact repeatability can be explored. Semi-automatic probe landing can be implemented.

Further, large signal measurement-based model generation using ANN will be a meaningful next step.

References

[1] P.A. 't Hart. "Cryogenic CMOS Characterization for Quantum Computer Applications". PhD thesis. TU Delft, 2022. DOI: 10.4233/uuid:0ab4ca6c-dc69-4207-970f-d3b9f0d9c5b4.

[2] "40 years of quantum computing". In: *Nature Reviews Physics* 4 (1 2022), p. 1. ISSN: 2522-5820. DOI: 10.1038/s42254-021-00410-6. URL: https://doi.org/10.1038/s42254-021-00410-6.

[3] Mohamed Aouad. "Compact modeling of FDSOI transistors at cryogenic temperatures". Theses. Université Grenoble Alpes [2020-....], July 2022. URL: https://theses.hal.science/tel-03852447.

[4] Imran Bashir et al. "A Mixed-Signal Control Core for a Fully Integrated Semiconductor Quantum Computer System-on-Chip". In: *ESSCIRC 2019 - IEEE 45th European Solid State Circuits Conference (ESSCIRC)*. 2019, pp. 125–128. DOI: 10.1109/ESSCIRC.2019.8902885.

[5] Arnout Beckers, Farzan Jazaeri, and Christian Enz. "Cryogenic MOS Transistor Model". In: *IEEE Transactions on Electron Devices* 65 (9 Sept. 2018), pp. 3617–3625. ISSN: 0018-9383. DOI: 10.1109/TED.2018.2854701. URL: https://ieeexplore.ieee.org/document/8424046/.

[6] Arnout Beckers et al. "Characterization and modeling of 28-nm FDSOI CMOS technology down to cryogenic temperatures". In: *Solid-State Electronics* 159 (2019). SI: EUROSOI 2018, pp. 106–115. ISSN: 0038-1101. DOI: https://doi.org/10.1016/j.sse.2019.03.033. URL: https://www.sciencedirect.com/science/article/pii/S0038110119301443.

[7] Arnout Lodewijk M Beckers. "Cryogenic MOSFET Modeling for Large-Scale Quantum Computing". In: (2021), p. 216. DOI: https://doi.org/10.5075/epfl-thesis-8365. URL: http://infoscience.epfl.ch/record/285710.

[8] Qing-Hong Bu et al. "Evaluation of a multi-line de-embedding technique for millimeter-wave CMOS circuit design". In: *2010 Asia-Pacific Microwave Conference*. 2010, pp. 1901–1904.

[9] Yang Cao et al. "A Complete Small-Signal MOSFET Model and Parameter Extraction Technique for Millimeter Wave Applications". In: *IEEE Journal of the Electron Devices Society* 7 (2019), pp. 398–403. ISSN: 2168-6734. DOI: 10.1109/JEDS.2019.2900202. URL: https://ieeexplore.ieee.org/document/8651585/.

[10] E. Catapano et al. "On the Zero Temperature Coefficient in Cryogenic FD-SOI MOSFETs". In: *IEEE Transactions on Electron Devices* 70 (3 Mar. 2023), pp. 845–849. ISSN: 0018-9383. DOI: 10.1109/TED.2022.3215097. URL: https://ieeexplore.ieee.org/document/9933050/.

[11] W. Chakraborty et al. "Cryogenic RF CMOS on 22nm FDSOI Platform with Record fT=495GHz and fMAX=497GHz". In: *2021 Symposium on VLSI Technology*. 2021, pp. 1–2.

[12] Wriddhi Chakraborty et al. "Characterization and Modeling of 22 nm FDSOI Cryogenic RF CMOS". In: *IEEE Journal on Exploratory Solid-State Computational Devices and Circuits* 7 (2 Dec. 2021), pp. 184–192. ISSN: 2329-9231. DOI: 10.1109/JXCDC.2021.3131144. URL: https://ieeexplore.ieee.org/document/9627653/.

[13] Maria-Anna Chalkiadaki. "Characterization and modeling of nanoscale MOSFET for ultra-low power RF IC design". In: (2016), p. 227. DOI: https://doi.org/10.5075/epfl-thesis-7030. URL: http://infoscience.epfl.ch/record/218534.

[14] Y. Cheng, M.J. Deen, and C.-H. Chen. "MOSFET Modeling for RF IC Design". In: *IEEE Transactions on Electron Devices* 52 (7 July 2005), pp. 1286–1303. ISSN: 0018-9383. DOI: 10.1109/TED.2005.850656.

[15] H. Cho and D.E. Burk. "A three-step method for the de-embedding of high-frequency S-parameter measurements". In: *IEEE Transactions on Electron Devices* 38 (6 June 1991), pp. 1371–1375. ISSN: 00189383. DOI: 10.1109/16.81628. URL: http://ieeexplore.ieee.org/document/81628/.

[16] J.P.G. van Dijk, E. Charbon, and F. Sebastiano. “The electronic interface for quantum processors”. In: *Microprocessors and Microsystems* 66 (Apr. 2019), pp. 90–101. ISSN: 01419331. DOI: 10.1016/j.micpro.2019.02.004.

[17] C. Enz. “An MOS transistor model for RF IC design valid in all regions of operation”. In: *IEEE Transactions on Microwave Theory and Techniques* 50 (1 2002), pp. 342–359. ISSN: 00189480. DOI: 10.1109/22.981286. URL: http://ieeexplore.ieee.org/document/981286/.

[18] C. Enz and Y. Cheng. “MOS transistor modeling for RF IC design”. In: *IEEE Journal of Solid-State Circuits* 35.2 (2000), pp. 186–201. DOI: 10.1109/4.823444.

[19] C. Esposito et al. “Extending the Open-Short de-embedding frequency via metal-l on-wafer calibration approaches”. In: IEEE, June 2022, pp. 1–4. ISBN: 978-1-6654-6894-7. DOI: 10.1109/ARFTG54656.2022.9896529. URL: https://ieeexplore.ieee.org/document/9896529/.

[20] A. Ferrero and U. Pisani. “Two-port network analyzer calibration using an unknown ’thru’”. In: *IEEE Microwave and Guided Wave Letters* 2.12 (1992), pp. 505–507. DOI: 10.1109/75.173410.

[21] G. Gonzalez. *Microwave Transistor Amplifiers: Analysis and Design*. Prentice Hall, 1997. ISBN: 9780132543354. URL: https://books.google.nl/books?id=-AVTAAAAMAAJ.

[22] Sumreti Gupta et al. “Cryogenic temperature DC-IV measurements and compact modeling of n-channel bulk FinFETs with 3–4 nm wide fins and 20 nm gate length for quantum computing applications”. In: *Solid-State Electronics* 185 (2021), p. 108089. ISSN: 0038-1101. DOI: https://doi.org/10.1016/j.sse.2021.108089. URL: https://www.sciencedirect.com/science/article/pii/S0038110121001349.

[23] Hung-Chi Han et al. “Cryogenic RF Characterization and Simple Modeling of a 22 nm FDSOI Technology”. In: *ESSDERC 2022 - IEEE 52nd European Solid-State Device Research Conference (ESSDERC)*. 2022, pp. 269–272. DOI: 10.1109/ESSDERC55479.2022.9947192.

[24] Pascal A. It Hart et al. “Artificial Neural Network Modelling for Cryo-CMOS Devices”. In: IEEE, Apr. 2021, pp. 1–4. ISBN: 978-1-7281-9306-9. DOI: 10.1109/WOLTE49037.2021.9555438. URL: https://ieeexplore.ieee.org/document/9555438/.

[25] R. M. Incandela et al. “Nanometer CMOS characterization and compact modeling at deep-cryogenic temperatures”. In: *2017 47th European Solid-State Device Research Conference (ESSDERC)*. 2017, pp. 58–61. DOI: 10.1109/ESSDERC.2017.8066591.

[26] Masaya Iwamoto et al. “Knowledge-based neural network (KBNN) modeling of HBT junction temperature and thermal resistance from electrical measurements”. In: *2017 IEEE MTT-S International Microwave Symposium (IMS)*. 2017, pp. 1069–1071. DOI: 10.1109/MWSYM.2017.8058777.

[27] Troels Emil Kolding. *A Four-Step Method for De-Embedding Gigahertz On-Wafer CMOS Measurements*. 2000.

[28] Roger B. Marks. “Formulations of the Basic Vector Network Analyzer Error Model including Switch-Terms”. In: *50th ARFTG Conference Digest*. Vol. 32. 1997, pp. 115–126. DOI: 10.1109/ARFTG.1997.327265.

[29] A.-S. Porret, J.-M. Sallese, and C.C. Enz. “A compact non-quasi-static extension of a charge-based MOS model”. In: *IEEE Transactions on Electron Devices* 48.8 (2001), pp. 1647–1654. DOI: 10.1109/16.936582.

[30] Douglas Kent Rytting. “Network Analyzer Accuracy Overview”. In: *58th ARFTG Conference Digest*. Vol. 40. 2001, pp. 1–13. DOI: 10.1109/ARFTG.2001.327486.

[31] Fabio Sebastiano et al. “Cryo-CMOS Electronic Control for Scalable Quantum Computing”. In: ACM, June 2017, pp. 1–6. ISBN: 9781450349277. DOI: 10.1145/3061639.3072948.

[32] Marco Spirito and Andrej Rumiantsev. *On-Wafer System Calibration for mm-Wave Frequency Applications, ARFTG/NIST Short Course*.

[33] Naoki TAKAYAMA et al. “A De-Embedding Method Using Different-Length Transmission Lines for mm-Wave CMOS Device Modeling”. In: *IEICE Transactions on Electronics* E93-C (6 2010), pp. 812–819. ISSN: 0916-8524. DOI: 10.1587/transele.E93.C.812. URL: http://www.jstage.jst.go.jp/article/transele/E93.C/6/E93.C_6_812/_article.

[34] Zhidong Tang et al. “Cryogenic CMOS RF Device Modeling for Scalable Quantum Computer Design”. In: *IEEE Journal of the Electron Devices Society* 10 (2022), pp. 532–539. ISSN: 2168-6734. DOI: 10.1109/JEDS.2022.3186979. URL: https://ieeexplore.ieee.org/document/9810355/.

[35] Yannis Tsividis. *Tsividis, Y. Operation and Modeling of the Mos Transistor, 2nd ed.; WCB/McGraw-Hill: Boston, 1998.*

[36] Ewout P Vandamme et al. *Improved Three-Step De-Embedding Method to Accurately Account for the Influence of Pad Parasitics in Silicon On-Wafer RF Test-Structures*. 2001.

[37] Zewei Wang et al. “Temperature-Driven Gate Geometry Effects in Nanoscale Cryogenic MOSFETs”. In: *IEEE Electron Device Letters* 41 (5 May 2020), pp. 661–664. ISSN: 0741-3106. DOI: 10.1109/LED.2020.2984280.

[38] Yunqiu Wu et al. “An Improved RF MOSFET Model Accounting Substrate Coupling Among Terminals”. In: *IEEE Microwave and Wireless Components Letters* 28.2 (2018), pp. 138–140. DOI: 10.1109/LMWC.2017.2780766.