Cryogenic CMOS LNA for RF readout of spin qubits

R. M. Incandela

Technische Universiteit Delft

TU Delft Delft University of Technology

Challenge the future

Cryogenic CMOS LNA for RF readout of spin qubits

by

R. M. Incandela

in partial fulfillment of the requirements for the degree of

Master of Science
in Electrical Engineering, Microelectronics

at the Delft University of Technology,
to be defended publicly on Friday October 28, 2016 at 13:30 PM.

Student number:	4419332	
Project duration:	August, 2015 – October, 2016	
Thesis committee:	Prof. dr. ir. E. Charbon,	TU Delft, supervisor
	Dr. F. Sebastiano,	TU Delft, daily supervisor
	Prof. dr. ir. K. A. A. Makinwa,	TU Delft

This thesis is confidential and cannot be made public until December 31, 2020.

TU Delft Delft University of Technology

Abstract

Quantum computation is bringing excitement and motivation into the scientific community to build a practical quantum computer. This would enable the solution of problems today intractable, thanks to the exponential decrease in the required number of operations with respect to a classical computer. Nevertheless, this extraordinary computer needs support from a classical computer to perform specific side tasks, such as quantum error correction, control and post-processing of the information where classical electronics is more efficient. Further, to enable the scalability of quantum computers reducing the huge amount of interconnections that are nowadays needed to perform quantum computation, classical electronics must be placed close to the quantum processor at cryogenic temperatures (4 K). To address this, firstly a compact model for simulation of CMOS at cryogenic temperature is proposed and secondly a CMOS Low-Noise Amplifier (LNA) is designed and tested at liquid Helium temperature (4 K). The cryoLNA is the first in this technology and will be employed in the RF electronic readout of spin qubit, replacing the existing discrete amplifier used in state-of-the-art experimental setup. The functionality of the CMOS LNA at deep cryogenic temperatures demonstrates the effectiveness of cryoCMOS and makes a first step towards the realization of integrated and scalable quantum computers.

Acknowledgement

The support I received during this whole year of work was crucial to accomplish this thesis.

I would first like to express my gratitude to my parents, for their limitless support during these past two years in Delft. They always provided me with a relaxed environment, especially in stressful periods, and shared their joy and pride for every success I obtained.

I am also very thankful to dr. Fabio Sebastiano for his supervision during the development of this thesis. I could not have asked for a better daily supervisor and I am glad I can continue working with him for my PhD.

I would like to thank my supervisor prof. Edoardo Charbon for giving me the opportunity to work in such a beautiful group which is the CoolGroup, and for having let me join this project for other four more years as a doctoral researcher.

I would like to thank Giulia, whose patience and love were always above everything, and my friend from Milan, who are coming to Delft to celebrate this important occasion with me.

Finally, 'Gli zii di Delft' also deserve my gratitude for the funniest time spent together, during these past two years.

Ad maiora!

R. M. Incandela
Delft, October 2016

Contents

List of Figures

List of Tables

1

Introduction

1.1. Quantum revolution

After the discovery of quantum mechanics more than a century ago, two quantum revolutions could be observed. The first one started with the discovery of the transistor and enabled the technological world we are living in now through the realization of pervasive electronic devices. In such devices, several quantum effects, such as tunneling and bandgap-based energy structures, are observed and exploited. The second revolution, which is blossoming right now, consists of exploiting other more fundamental quantum properties, such as superposition and entanglement, to build a more powerful computing framework, that would enable solving problems that are nowadays intractable by classical computers. Already in 1982 [1], Feynman suggested to implement a new type of computer exploiting quantum effects to efficiently simulate quantum systems. A quantum computer, with its quantum bits (qubits), will be able to exploit superposition and entanglement to exponentially decrease the amount of operations that a classical computer would perform for the same task. The big advantage can be understood by the following simplistic example: a combination of two classical bits can only assume four values, '00', '01', '10' or '11' while two qubits in superposition can have infinite combinations described by $\alpha|00\rangle + \beta|01\rangle + \gamma|10\rangle + \delta|11\rangle$ where α, β, γ and δ represent the probability of finding the two qubits in state $|00\rangle$, $|01\rangle$, $|10\rangle$ or $|11\rangle$, respectively. Hence, to discriminate among the four classical combinations only two numbers are needed, which are the bits, while for a two-qubits state four numbers are requested: α, β, γ and δ. The same for three classical bits where only three numbers are needed while eight numbers are needed for three qubits in superposition. In other words, for N classical bits, N pieces of information are available while for N qubits, 2^N numbers are provided. Rephrasing it, an N-qubits quantum computer can be said to be equivalent to a 2^N classical computer. Thanks to this huge exponential improvement adn to the inner probabilistic nature of quantum mechanics, quantum computers could address many relevant problems among which [2]:

- Random number generator: it is impossible, nowadays, to make a fully random number generator in a classical computer because of the intrinsic deterministic nature of classical physics. Being quantum physics completely probabilistic this issue is solved. In fact, a qubit can be in a superposition of 0 and 1 state at the same time. Once it is measured, it will collapse to one of these two states, with a probability of 50 %, which is a fully random process.

- Quantum cryptography: exploiting entanglement yields completely secure cryptographic keys. If two people want to exchange secret information, each of them can be provided with one qubit of an entangled pair. When they measure their respective qubit, a random number will be obtained that can be used as cryptographic key for communication. The real breakthrough is that, thanks to quantum entanglement, this random number, and hence the key, will be the same for both parties without possibility of cloning, enabling safe and secure communication.

- Quantum simulation of molecules: simulations of physical processes that are now computationally impossible, could be completed in a couple of days with a quantum computer. This will enable improving the efficiency of many industrial processes that are not fully optimized due to lack of the

understanding that can be gained through simulations. The improved efficiency of processes such as nitrogen fixation in fertilized will result in huge worldwide energy savings, with a significant impact on the solution of big environmental problems like global warming.

- Discovery of new materials: by enabling the quantum simulation of molecules, new material properties or even new combinations of materials can be found, such as high-temperature superconductor that could enable transmission of power without losses.

Scientists have tried hard to realize a quantum computer but the most advanced quantum processor in 2016 consists of only 9 superconducting qubits [3], while the minimum number of logical qubits needed to perform the simplest task should be around 100, accomplished with around one hundred thousand physical qubits. Nevertheless, excitement and big investments are boosting quantum computing research, thus leading researchers to believe that a practical quantum computer will become available in the near future.

1.2. The need for classical electronics

Although a quantum processor can overtake the potentiality of the best classical supercomputer, it must be noticed that classical computers will remain the cheapest solutions for many common tasks. Svore, in [2], envisions a future where quantum computers are placed in big datacenters and used as a coprocessor to support classical computers. Only few qubits will be implemented locally in portable devices to enable secure cryptography and exchange of secret pieces of information.

Another reason of having a classical and a quantum processor side by side is the so-called quantum error correction(QEC). In 2000, DiVincenzo [4] proposed 5 main criteria that a practical quantum computer must have. The third states that a qubit must be robust against environment harshness for a time longer than the single operation time. This gave rise to the QEC, which is a framework of classical electronics and algorithms to retrieve the correct information if the qubit state is lost.

Furthermore, from a pragmatic point of view, a quantum computer will be practical only if it is small, cheap and fast enough [5], otherwise if the technological costs of implementation overcome the benefits, no advantages will be brought to the society.

From this introduction it is clear that classical computer support is still needed for a practical quantum computer to be feasible. In Fig. 1.1, [5] shows a possible implementation of a quantum system, where a classical computer postprocesses the data, while another classical processor is placed close to its quantum counterpart mainly for QEC.

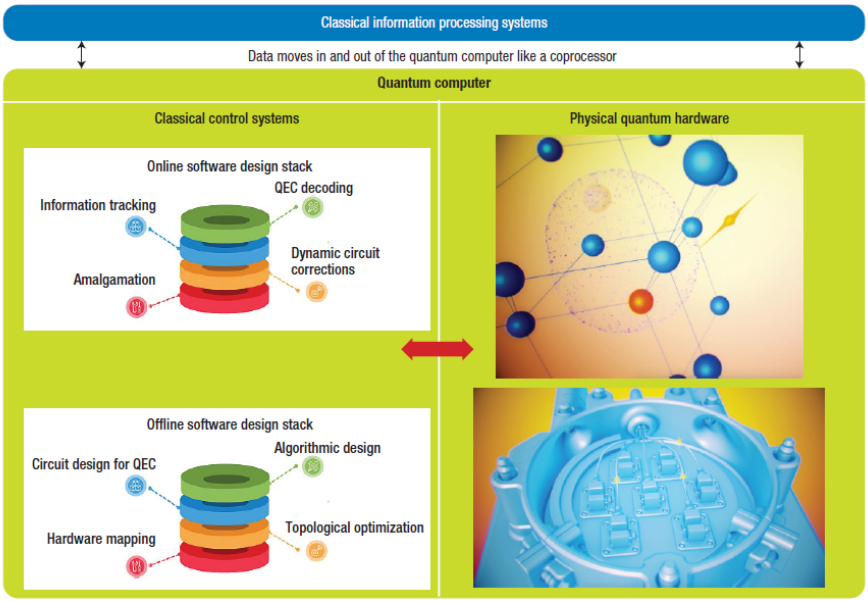

Figure 1.1: Envisioned quantum computer architecture where two classical processors support the quantum computation. [5]

1.3. The need for cryoCMOS

Since classical control electronics and quantum processors will live together, the same practical ingredients needed for the quantum processor hold for its classical counterpart. The adjectives small, cheap and fast enough immediately translates into CMOS technology. For many years already, we are benefiting from CMOS processors and all classical computers are basically based on advanced CMOS technologies (see Moore's law). This technology has enabled the the scalability of billions of bits of information in chips large only a couple of millimiters squared and limited cost. The same is required for quantum computers.

Nevertheless, unlike classical bits, state-of-the-art qubits can only operate at extremely low temperatures ($\approx$ mK range) [6]. Large dilution refrigerators (Fig. 1.2), are nowadays used to keep the quantum devices at deep cryogenic temperatures and long interconnections are routed from the deep cryogenic chamber to room-temperature bench-top instruments that read out and control the quantum bits. In a practical quantum computer this would never work because billion physical qubits would require as many interconnections. By recalling Fig. 1.1 and imagining a chip where both classical and quantum processors are placed together, this means that classical electronics must operate at deep-cryogenic temperatures.

Figure 1.2: Dilution refrigerator. Each stage is at a different temperature, starting from mK range up to tens of K. Courtesy of Oxford Instrument.

1.4. Thesis objective

The main objective of this thesis is to prove the functionality of a CMOS complex circuit at 4 K. A Low-Noise Amplifier (LNA) will be designed to be the first block in the readout chain of spin qubits. Spin qubit is one possible physical implementation of qubit based on the spin of electrons. State-of-the-art experiments use off-the-shelf bulky components to read out and control spin qubits. This design will be a first step towards a cheap and scalable quantum computer, where CMOS will be the core technology. A preliminary objective is to study and model the behavior of CMOS devices at cryogenic temperatures to build cryogenic compact models that can be employed in traditional IC design tools for the design of the proposed LNA.

1.5. Thesis outline

The thesis is organized as follows:

Chapter 2 gives an overview of spin qubits and how readout can be performed. The LNA specifications will be derived.

Chapter 3 presents the behavior of CMOS at cryogenic temperatures and, based on this, the derivation of a compact model will be shown.

Chapter 4 includes the design of the Low-Noise Amplifier while chapter 5 presents the measurements and tests of the LNA.

Chapter 6 summarizes the conclusions and presents how future works can extend the results of this thesis.

2

Readout

In this chapter, an overview of spin qubit and their readout is firstly given. Then, specifications for the readout developed in this thesis will be derived.

2.1. Spin Qubits

Qubits are the quantum counterpart of classical bits. There are different ways of implementing a qubit [7], among which the most promising are: superconducting qubits, spin qubits, topological qubits and diamond-based qubits. This thesis focuses on the readout of spin qubits and, therefore, a general overview of only this implementation will be given.

In spin qubits, quantum information is stored in the spin of electrons. In order to manipulate a single qubit, a single electron needs to be isolated from the external world and electrically controlled. To do so, particular structures, called quantum dots, are fabricated. In Fig. 2.1, a sketch of a realistic implementation of spin qubits is shown. This structure can be implemented in GaAs substrate, as well

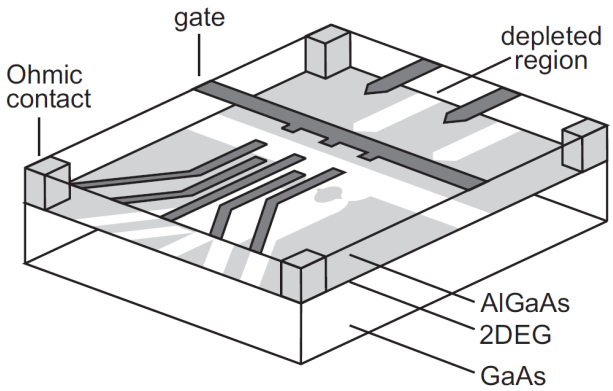

Figure 2.1: Quantum dot. [8]

as SiGe and Si. In the proposed example in Fig 2.1, a layer of AlGaAs is deposited on top of a GaAs substrate. Due to difference in bandgap of the two materials, a 2-D electron gas (2DEG) forms at their interface [9]. By applying negative voltages on these gates, it is possible to deplete the region beneath (white parts) and, therefore, form ad-hoc regions to trap electrons. If such regions are made small enough, a quantum dot (QD) is formed. In Fig. 2.1, the white strips in the 2DEG are depleted regions, while the grey ones are filled by electrons. In the center a quantum dot is shown: a very tiny circled region. By proper tuning of the voltages, a single electron, from the large reservoirs all around, can be transferred to the dot and trapped there, as shown in Fig. 2.2. Succesively, the spin of the electron in the quantum dot can be controlled by the voltages surrounding the gates and read out. Being the

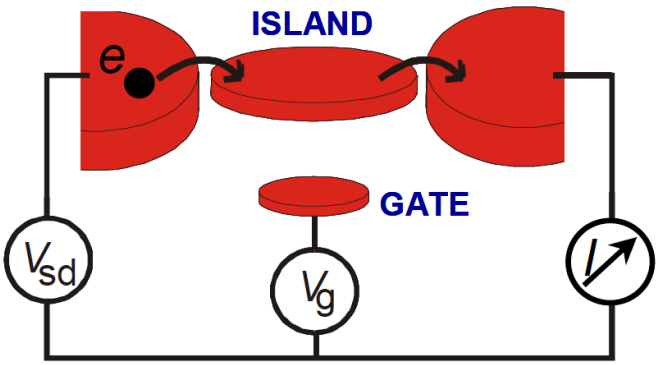

Figure 2.2: Sketch of a quantum dot from the potentials point of view. Here the island is the quantum dot and the two regions aside are the reservoirs. V_{sd} is the potential applied at the gate around the dot to tune the tunneling barrier, while V_g is the potential applied to the gate aligned with the QD used to tune the electrochemical potential of the dot. Figure adapted from [7]

quantum dot extremely small, the energy levels that an electron can have are quantized as inside an atom. In the lowest energy level, according to Pauli's principle, two electrons can be placed, one with spin up and one with spin down. In order to create a two-level system, this energy level can be further splitted in order to make two sub-levels that can accommodate a single electron: one sub-level will host a spin-down electron and the other sublevel a spin-up electron. This is done by applying a strong magnetic field and exploiting the Zeeman effect (Fig. 2.3). Moreover, it is required that the thermal energy of the electron is lower than the Zeeman splitting ΔE, otherwise the electron could jump up and down into different energy levels without any control. For this reason, apin ubits are usually operated at temperatures of few tens of mK.

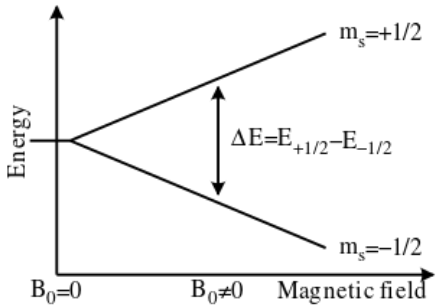

Figure 2.3: Effect of applying a magnetic field to a quantum dot: splitting of the energy levels into sub-levels. Picture taken from [10]

2.2. Readout

2.2.1. Spin-to-charge conversion

Since it is difficult to directly read the spin of electrons, scientists developed a method to convert spin information into charge information, thus enabling a much simpler electronic readout. In Fig. 2.4b an example of a quantum dot and its sensor is shown. The quantum dot is the circled region among the gates M, P, R and T. By tuning the voltages at M,R and T, the tunneling barrier, from the dot to the reservoir, can be varied such that if the electron has spin up it will not leave the dot, otherwise it will (Fig. 2.4a). This translates the direction of the spin into charge variation inside the dot. By capacitively coupling the potential of the dot to an external charge sensor, the spin can be read out. The quantum point contact (QPC) and the charge sensing dot (CSD) are the most used sensors to detect this charge variation inside the dot [7].

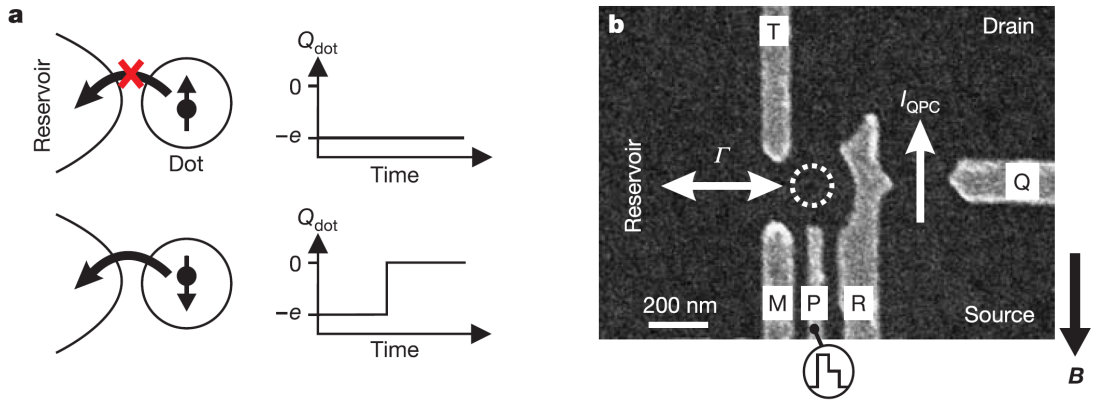

Figure 2.4: a) shows the the effect of the spin onto the charge in the dot while b) shows the quantum dot along with its respective QPC [11] and gates

2.2.2. QPC

On the right side of gate R in Fig. 2.4, a very narrow channel, the QPC, is made. The conductivity of such a channel is tuned by a voltage applied at gate Q. Being the channel extremely narrow, its conductance is quantized and can only assume values that are multiple of the quantum conductance, i.e. $G_q = 2e^2/h \approx 1/\ 13\ k\Omega$, where e is the charge of the electron and h is the Planck's constant. The conductance of the channel, as a function of the voltage applied at gate Q is shown in Fig 2.5.

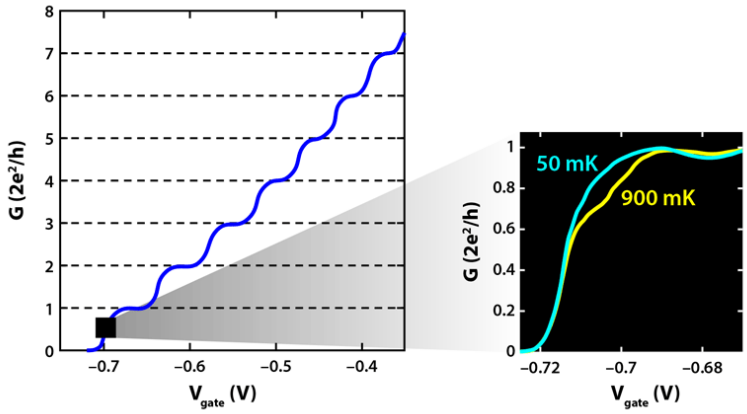

Figure 2.5: Quantum conductance as function of V_{QPC} [12]

Thanks to the shape of the conductance, the QPC is an extremely sensitive charge sensor: if biased in the steepest slope between two plateaus, a small variation on V_{gate} produces a large impedance change that can be then read out. Furthermore, the steepness of the conductance strongly depends on the temperature this sensor is operated at and degrades if it is placed at high temperature, as shown in the inset of Fig. 2.5. The mathematical description of these structures lies beyond the scope of this thesis, nevertheless, it can be qualitatively said that the most sensitive part is the first step between pinch-off and first plateau where $G = 0.7G_q$. This will be important for the derivation of the electrical specifications.

2.2.3. CSD

By placing another quantum dot close to the main dot, an alternative sensor is formed. The principle is very similar to the QPC: by capacitive couple the two dot potentials, a spin variation in the main dot (and consequently the electron tunneling) produces a potential variation in the second dot. By placing two reservoirs on each side of the second dot (same principle as in Fig. 2.2), single-electron

Table 2.1: Electrical model summary of the QPC

R_{QPC}	ΔR/R	Bandwidth	Noise Temperature
25 kΩ	1 %	10 MHz	1 K

current can flow. In order for the current to flow, the potential of the dot must be set as in Fig. 2.6b, between the potential of the two reservoirs. On the other hand, if the potential of the dot is like in Fig. 2.6a, such that only one energy level is below the reservoirs' potentials, then electrons can not flow and Coulomb blockade is observed. In Fig. 2.6c, the current as a function of the potential of the dot is shown: where the current is zero, Coulomb blockade is happening. If the dot potential is set in between these two regions, current-on and current-off, and capacitively coupled to the main dot potential, a very sensitive detector is created.

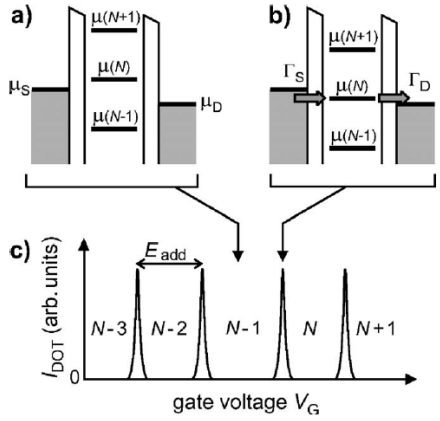

Figure 2.6: Current peaks due to tunneling of one electron versus the voltage applied to the gate (Fig. 2.1), [7].

This thesis will focus on the spin readout through a Quantum Point Contact. From the electrical point of view, a QPC can be modeled as a varying resistance whose nominal value is 25 kΩ. We will target the specifications of the experimental setup used in the group of Prof. Vandersypen at TU Delft. For this setup, the variation of this resistance is 1 % of the nominal value and has a duration in the order of 100 ns, resulting in a required detection bandwidth of 10 MHz. In terms of noise, a quantum point contact has an equivalent noise temperature of around 1 K for the 'Delft' QPC. In tab. 2.1 the final electrical model is summarized.

2.2.4. Impedance Readout

One method to read out the impedance of the QPC when the spin changes consists in measuring the reflection coefficient in a transmission line terminated by the Quantum Point Contact. This is done with the radio-frequency reflectometry setup shown in Fig. 2.7. The nominal impedance of the QPC, i.e. the impedance when the electron is in the quantum dot, is matched (through a frequency selective matching network) to a 50-Ω characteristic impedance. A signal generator at room temperature (RT) generates an RF tone with a relatively large amplitude. Such tone is attenuated at 4 K, in order to reduce the thermal noise from the RT generator while adding a low noise in the attenuator thanks to reduced operating temperature of the attenuator itself. The signal reaches the QPC and if the electron has spin up and, hence, stays in the QD, all the power is absorbed by the QPC and none is reflected back. Otherwise, if the spin is down, the electron leaves the dot, the QPC impedance change so that the impedance at the line end will not be matched anymore and a fraction of the injected power is reflected back. The reflected power is routed through a directional coupler and then amplified by the cascade of a cryogenic amplifier (SiGe discrete amplifier [13]) and a RT amplifier. The signal at RF is then mixed with the input signal to implement an homodyne demodulation to baseband. Any high-frequency component at the mixer output is filtered out and a baseband output is produced for further processing.

Compared to other readout alternatives, such as DC current readout through a transimpedance

amplifier [14], this method is not affected by the parasitic capacitances of the long wires from RT to Liquid Helium Temperature (LHT) that may limit the detection bandwidth. The only limitation in detection bandwidth is the quality factor Q of the matching network. A very important disadvantage, on the other hand, is the limited scalability offered by this method because of the need of a matching network and a directional coupler that cannot easily be integrated.

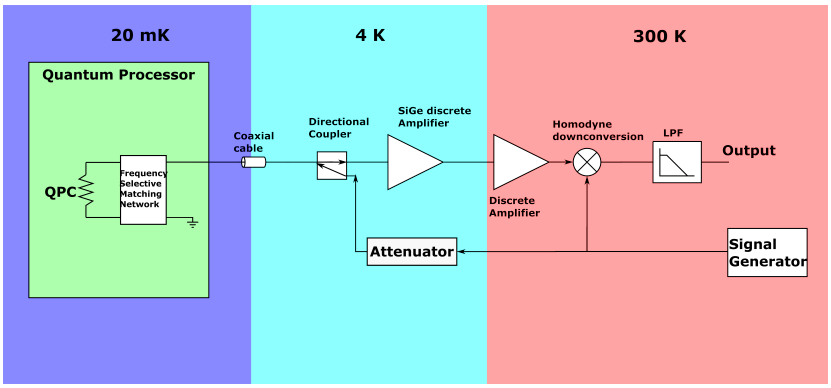

Figure 2.7: RF reflectometry setup

2.3. LNA Specifications

In the future, both qubits and electronics will be placed onto the same chip at cryogenic temperatures. Therefore, electronics must be designed for very low temperatures and must be able to perform correctly. The goal of the thesis is to make a first step towards CMOS integration of the RF-reflectometry setup to boost the scalability of quantum computers. To achieve this goal, a CMOS Low-Noise Amplifier (LNA) is designed, which is the first critical block in the readout chain. In this section, the requirements for the Low-Noise Amplifier will be derived.

2.3.1. Signal

The typical matching network used in RF reflectometry setup is the LC network shown in Fig. 2.8. L

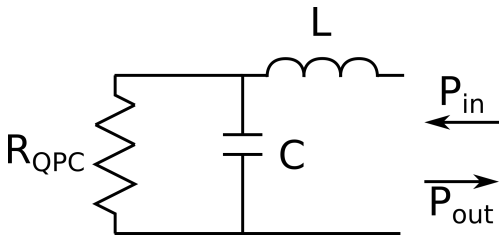

Figure 2.8: LC Matching Network

is an off-the-shelf component while the capacitor C accounts for the parasitics of the QPC and of the inductor package. Typical values for this component, in the 'Delft' setup, are C = 736 fF and L = 921 nH achieving a resonating frequency f_0 = 193.4 MHz. From the L-matching network equations:

$$Q = \sqrt{\frac{R_{QPC}}{R_s} - 1} = 22.3 \tag{2.1}$$

where R_{QPC} is 25 kΩ and R_s = 50 Ω. From the quality factor, the bandwidth of the resonator can be retrieved:

$$BW = \frac{f_0}{Q} = 8.6MHz \tag{2.2}$$

that is smaller than the required bandwidth of 10 MHz. The reflection coefficient is defined as:

$$\Gamma_L = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{2.3}$$

where Z_L is the impedance seen by looking through the matching network towards R_{QPC} and Z_0 is 50 Ω. At the resonance frequency ($f_0 = \frac{1}{\sqrt{LC}}$), the reflection coefficient can be rewritten as:

$$\Gamma_L = \frac{\frac{L}{R_{QPC}C} - Z_0}{\frac{L}{R_{QPC}C} + Z_0} = \frac{L - R_{QPC}CZ_0}{L + R_{QPC}CZ_0} \tag{2.4}$$

To a first order approximation, the variation of Γ_L as function of ΔR_{QPC} can be written as:

$$|\Delta\Gamma_L| = \frac{2LCZ_0}{(L + RCZ_0)^2}\Delta R_{QPC} \tag{2.5}$$

With values of L and C stated above and with $\Delta R_{QPC}/R_{QPC}$ = 1% , this amounts to $\Delta\Gamma_L/\Gamma_L$ 0.5 %. The power reflected is defined as:

$$P_{reflected} = P_{incident} \cdot |\Gamma_L|^2 \tag{2.6}$$

Inserting the numbers above in the equation yields a reflected power equal to 0.0025 % of the incident power.

This, in reality, is an upper bound: in fact, some of the signal will be dissipated in the matching network and some will be reflected back. Those were neglected in this simplified analysis. In general, qubits can tolerate very low level of powers, around -90 dBm [15]. In this scenario, the reflected power is less than -135 dBm.

2.3.2. Input matching of the amplifier

In order to absorb the power reflected, the amplifier itself needs to be matched to the 50-Ωimpedance of the line. To quantify the quality of the input matching s11 = $V_{reflected}/V_{incident}$ needs to be specified. Considering that the input power is already really small, ideally s11 should be really small. Moreover, this would help in reducing possible multiple reflections and/or standing waves that could hamper both the readout and the state of the qubit. For this reason, an s11 < -10 dB was chosen as a requirement, enabling correct readout by reflecting only one tenth of the incoming power.

2.3.3. Noise

State-of-the-art reflectometry setups are mainly limited by noise. The main sources of noise are [16]:

- Shot noise of QPC
- Thermal noise of the matching network
- Thermal noise of the cryogenic amplifier

The assumed values for this three contributors are T_{noise} = 3 K for the cryogenic amplifier [13], T_{noise} = 1 K for the QPC and T_{noise} = 5 K for the matching network [16], summing up to an equivalent noise temperature of 9 K.

2.3.4. SNR

From the numbers above (signal, noise and bandwidth), the SNR is less than 1 (SNR = -135 dBm - (-119 dBm [1]) = -16 dB). For this reason, post-processing of data is needed nowadays, to reduce the bandwidth of the noise and signal in the kHz range [15]. To estimate how much SNR would be needed to properly readout the qubit state, it is possible to think of the reflectometry readout as an On-Off-Keying (OOK) demodulation. From [17], a bit error rate (BER) below 10^{-14} is desirable to run a practical quantum algorithm such as the famous Shor's algorithm. For OOK and for BER < 10^{-14}, an SNR larger than 20 dB is necessary. This is clearly very far from the performance of state-of-the-art readouts. To reach the target, the first improvements must come from the charge sensor: sensitivity must be enhanced and noise reduced. It was then decided that the LNA should be able to reach the same noise level as the state-of-the-art discrete cryogenic amplifier shown in Fig. 2.7. This would enable a direct replacement of the discrete SiGe amplifier with the proposed LNA, without, in principle, altering the performance of the system.

[1] $P_{noise} = kT_{noise}B$, where B = 10 MHz and T_{noise} = 10 K.

From the datasheet of the amplifier [13], a noise temperature of 3.4 K is reported up to 1 GHz, with the amplifier measured at 23 K. Assuming that noise power scales proportionally with absolute temperature, this translates into a noise temperature of 0.6 K for the SiGe amplifier that operate at 4 K in the 'Delft' setup 2.7. This is computed according to:

$$T_{4K} = T_{23K} \cdot \frac{4}{23} = 0.6K \tag{2.7}$$

This, in turn, corresponds to a Noise Figure (NF) of 0.009 dB according to:

$$NF_{dB} = 10log_{10}\left(1 + \frac{T_{4K}}{T_{ref}}\right) \tag{2.8}$$

where T_{ref} = 290 K. Such a low NF sets the input-referred noise of the amplifier to less than $40\frac{pV}{\sqrt{Hz}}$ from:

$$NF = 1 + \frac{S_{in}}{4kT_{ref}R_s} \Rightarrow S_{in} = (NF - 1) \cdot 4kT_{ref}R_s \tag{2.9}$$

where S_{in} is the power spectral density of the amplifier, referred to the source and R_s is 50 Ω.

2.3.5. Power

In conclusion, a requirement on power consumption is set by the specifications of current dilution refrigerators that, at 4 K, can, approximately, offer 1 W of cooling power. In the future, when quantum processors with a much larger number of qubits will be fabricated, power dissipation required by the electronics will be a limiting factor. Our target is to implement a readout that can be used for 1000 qubits without exceeding the power budget of the dilution refrigerator at 4 K, i.e. dissipating ≈ 1 mW per qubit read-out. One possibility to achieve this would be frequency multiplexing of qubits, presented for the first time in [18]. The idea consists of having different matching network for each qubit, all tuned at different frequencies. By then sending a train of pulses, at all qubit frequencies, only the qubits with spin down will reflect power and therefore only tones at those frequency will be read back. This approach requires a wideband amplifier, able to amplify all those frequencies with the same gain. By sharing the same amplifier for several qubits, the equivalent power per qubit reduces proportionally. For this reason, the designed LNA should be compatible with frequency multiplexing and hence have a large bandwidth. Power per qubit is, then, the figure of merit the LNA will be optimized for. In the following chapters, the bandwidth of the qubit will be assumed to be equal to 10 MHz and, inferring some spacing between the qubit channels, the total bandwidth required per qubit is assumed equal to 20 MHz.

2.3.6. Summary

In summary, a cryo-CMOS LNA will be designed, to be used in the RF-reflectometry readout, with a noise performance comparable to the existing discrete SiGe amplifier. This low level of noise requires large power consumption. For this reason, the LNA should be compatible with a frequency-multiplexing setup that requires a large bandwidth to allocate many qubits, thereby reducing the equivalent power consumption per qubit. Each qubit channel will be assumed to require 20 MHz bandwidth. To assess the performance of the LNA, we use the following figure of merit:

$$\frac{P}{qubit} = \frac{P_{LNA}}{BW_{LNA}} \cdot BW_{qubit} = \frac{P_{LNA}}{BW_{LNA}} \cdot 20MHz \tag{2.10}$$

The valid bandwidth of the LNA will be considered to start from 200 MHz, which is the frequency at which the existing qubit is placed now (193.4 MHz). Therefore, the denominator can be rewritten as BW_{max} - 200 MHz.

Although a power consumption in the order of 1 mW/qubit would enable the readout of approximately 1000 qubits in existing refrigerators, it was unclear at the beginning of the project whether this could be achieved. Consequently, while P/qubit = 1 mW would be the ultimate goal, for this design, P/qubit will be minimized. In Fig. 2.9 the envisioned architecture is shown, to be compared with 2.7. In Table 2.2, a summary of the final top-level specifications for the LNA can be found:

Moreover, this will be the first CMOS block in this technology to be interfaced with spin qubits at 4 K. Apart from the specific application, it will be a proof of concept of the low-noise and high-frequency capabilities of cryoCMOS.

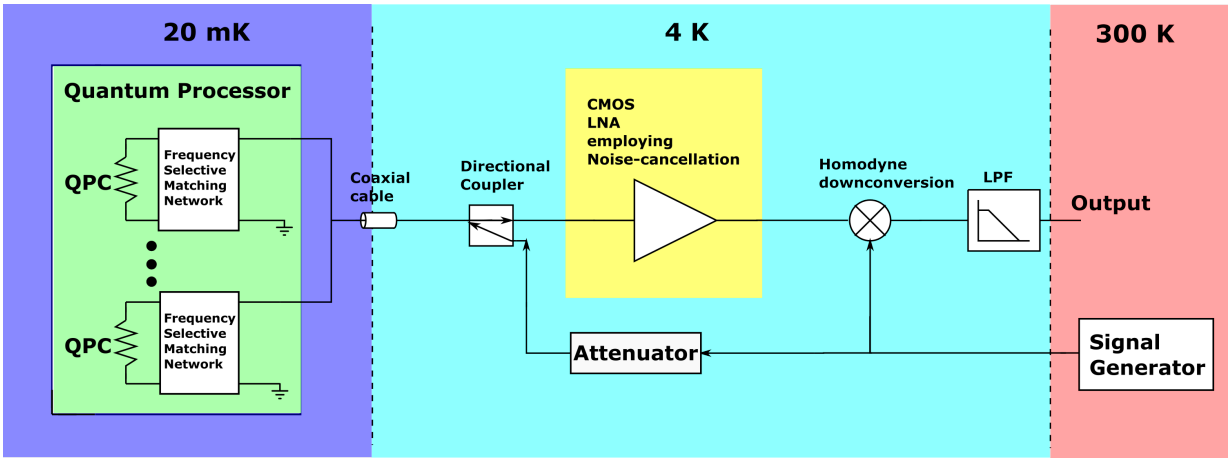

Figure 2.9: Envisioned RF-reflectometry setup, including the noise-cancelling LNA and many qubits frequency multiplexed. Also the other electronics components were placed at 4 K.

Table 2.2: Final top-level specifications

Operating temperature	Noise Figure	Input noise	Input Impedance	$S11_{max}$	Bandwidth	Power/qubit
4 K	0.009 dB @ T = 4 K	40 $pV/\sqrt{Hz}$	50 Ω	-10 dB	Largest achievable	Minimum

3

Cryogenic CMOS modelling

In this chapter, the need of CMOS models at cryogenic temperature will be highlighted. Moreover, an algorithmic procedure of how to adapt room temperature model to cryogenic temperatures will be presented, based on the physics and the understanding of cryogenic CMOS anomalies described in the next section. The new derived cryogenic model will be used for the design of the Low-Noise amplifier presented in the next chapter.

3.1. Cryogenic behavior and anomalies of CMOS transistors

The physics of Silicon devices at low temperature is very well understood nowadays. Many experiments have been performed and solid theories developed. From solid-state physics, if Silicon is placed at low temperature, important changes in carrier concentration and mobility are firstly observed [19]. In particular, at temperatures below 40 K, carrier concentration decreases because of freeze-out effects while mobility increases because of reduced scattering events [20]. These concepts can also be applied to CMOS transistors. Although one would expect that substrate freeze-out would stop the conduction, free carriers are still present in the degenerately doped source and drain regions, and can be attracted to create the channel below the gate. Once the channel is formed, the low temperature brings many benefits to cryoCMOS circuits: higher switching speed, steeper subthreshold slope, higher transconductance. These advantages have motivated the scientific community to deeply study the behavior of CMOS at low temperatures and, in this section, a detailed overview of CryoCMOS behavior and anomalies will be given, mainly focused on the technology used in the design described in chapter 4 (SSMC 0.16 µm CMOS).

3.1.1. Mobility increase

As mentioned before, thanks to the reduced scattering mechanisms at lower temperature, a net increase in mobility is observed. It is well-known that mobility is affected by different scattering mechanisms, which are strongly dependent on temperature [20]:

- ↑ **Phonon scattering** happens because of electrons hitting the vibrating silicon lattice. Because of the atomic scale, vibrations are quantized in quanta called phonons. All existing models agree that, as temperature decreases, mobility increases (as shown in the small arrow) because vibrations in the lattice decrease and, consequently, less phonons with whom to collide.
- ↑ **Velocity saturation**: at high lateral electric fields, the velocity of the carriers saturates and, consequently, mobility is also reduced. At low temperatures, velocity saturates at higher electric fields and, therefore, mobility increases.
- ↓ **Surface scattering**: because of manufacturing imperfections, the silicon surface of the oxide is not perfect. Hence, carriers scatter, degrading the mobility. If temperature decreases, there will be more chance of electron-surface collision since carriers will have less kinetic energy[1]

[1]Here all the electrical potentials are assumed to be constant over temperature and only thermal energy changes.

- ↓ **Ionized impurity scattering**, also called Coulomb scattering, happens because of high level of doping and scattering of carriers due to dopants. For the same reason as surface scattering, if the temperature decreases, the mobility decreases as well because the chance of the carriers being scattered augments.

In conclusion, combining these scattering mechanisms with Mathiessen's rule, the equivalent mobility increases as temperature decreases, until Coulomb and surface scattering become dominant and mobility saturates, as it is shown in Fig. 3.1.

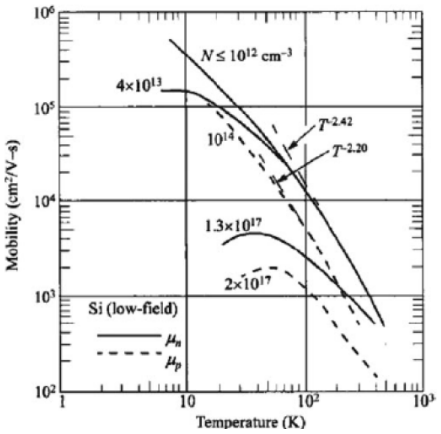

Figure 3.1: Mobility versus temperature for different doping levels, for NMOS and PMOS devices [21]

3.1.2. Threshold voltage

Threshold voltage is defined as the gate voltage at which the inversion layer below the gate is formed and drift-dominated conduction can start. This parameter strongly depends on temperature: in fact, the formation of the channel also relies on the kinetic energy of carriers. If carriers do not have enough thermal energy, more potential energy is necessary to compensate for it and make the carriers reach the region below the gate from the substrate. The behavior of the threshold voltage versus temperature is shown in Fig. 3.2 for different sizes of MOS transistors, showing that the trend is inversely proportional to temperature.

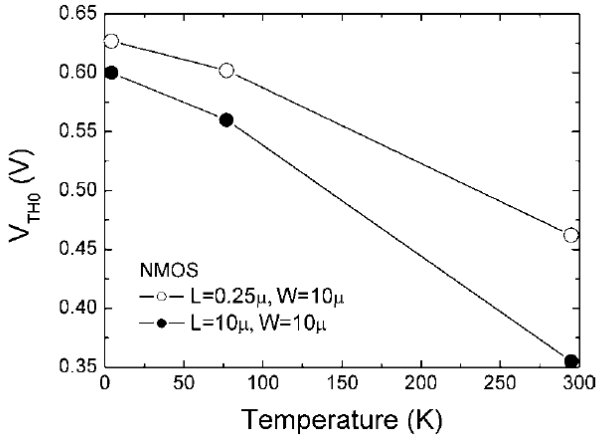

Figure 3.2: Threshold voltage versus temperature for different size of NMOS transistors [22]

3.1.3. Subthreshold slope

Subthreshold slope is defined as the amount of gate-to-source voltage needed to change the subthreshold current by a decade [23]. The subthreshold current, including all the dependences of temperature

can be found in [20] and reported here for simplicity:

$$I_{subth} = \left(\frac{\mu_0 W_{eff}}{2L_{eff}}\right)\left(\frac{kT}{q}\right)^2 \sqrt{\frac{q\epsilon_{Si}N_B}{\phi_B}} \cdot e^{\left[\frac{q(V_{GS}-V_T)}{nkT}\right]}\left[1 - e^{-\frac{qV_{DS}}{nkT}}\right] \tag{3.1}$$

where μ_0 is the mobility, W_{eff} and L_{eff} are the sizes of the transistor, k is the Boltzmann's constant, T is temperature, q is the electron charge, $n = \frac{C_{ox}+C_{j0}}{C_{ox}}$, C_{ox} is the oxide capacitance, C_{j0} is the channel-to-bulk capacitance, N_B is the bulk doping, ϕ_B is the built-in potential, V_{GS} is the gate-to-source voltage and V_{DS} is the drain-to-source voltage. If the V_{DS} dependence is discarded and the derivative with respect to V_{GS} is taken and then inverted, the subthreshold slope can be found:

$$SS = \left(\frac{dI_D}{dV_{GS}}\right)^{-1} = ln(10)\frac{kTn}{q} \tag{3.2}$$

Based on Eq. 3.2, one would expect a linear variation with temperature but the dependence of n on C_{j0} makes the temperature dependence more complicated. A plot of subthreshold slope as function of temperature is shown in Fig. 3.3, showing a sublinear increase in slope. A steeper subthreshold slope

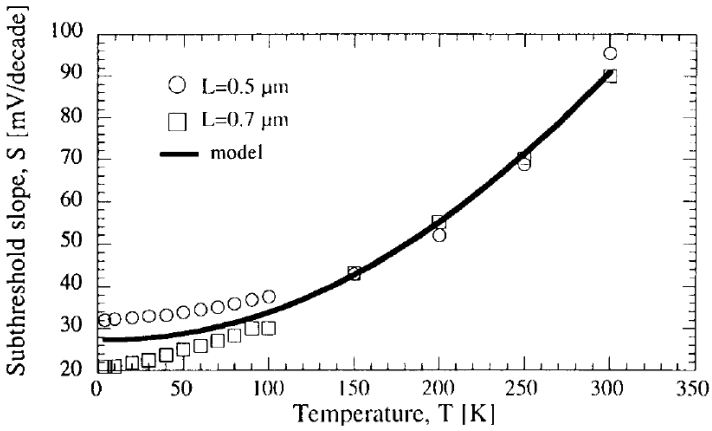

Figure 3.3: Subthreshold slope versus temperature. Symbols are measurements for two different CMOS technologies, while the thick line represents the model based on Eq. 3.1 from [20]

would allow faster switching and a huge benefit for digital circuits.

3.1.4. Kink effect

A plot of I_D versus V_{DS} is shown in Fig. 3.4, from [24], for two different sizes of NMOS transistors, with substrate doping $N_a = 1.5 * 10^{15}$ cm^{-3} and oxide thickness of 120 nm. At $V_{DS} \approx Vdd/2$ a big increase in current is observable. This jump in current is called 'kink'. The kink effect was observed in older technologies ([25], [26]). The most accepted theory on kink effect is from [27] and states that the kink is mainly caused by the combination of impact ionization and freeze-out effects. In fact, due to freeze-out of the substrate, the impedance of the latter increases dramatically. At the same time, impact ionization due to high lateral electric field, generates carriers that have to flow out through the substrate. This leakage current across a very high-impedance substrate creates a large voltage drop across the bulk, decreasing abruptly the threshold voltage and therefore producing a jump in the bias current. Furthermore, after the jump, the current saturates again: this is due to the parasitic diode between source and bulk. In fact, the voltage drop on the bulk is the same as the voltage across this diode (if the source is at ground): if the voltage increases above the forward voltage of the p-n junction, the diode turns on and fixes the voltage at the bulk, sinking all the extra-current. In Fig. 3.5, a cross-section of an NMOS transistor is shown, depicting the main causes of the kink: impact-ionization current across a high impedance, increasing the internal potential of the body, that is pinned by the diode. Fortunately, kink is found only in older technologies: CMOS technologies with feature size below 0.16 μm have not shown any kink ([22], [29]). The reason of this is due to higher doping

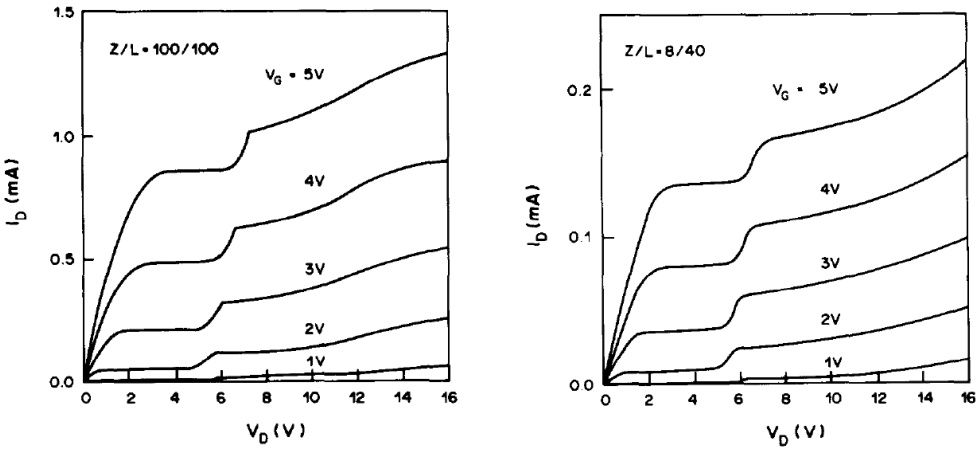

Figure 3.4: I_D-V_{DS} curve for two different NMOS transistors. W = 100 µm, L = 100 µm on the left, W = 8 µm, L = 40 µm on the right [24]

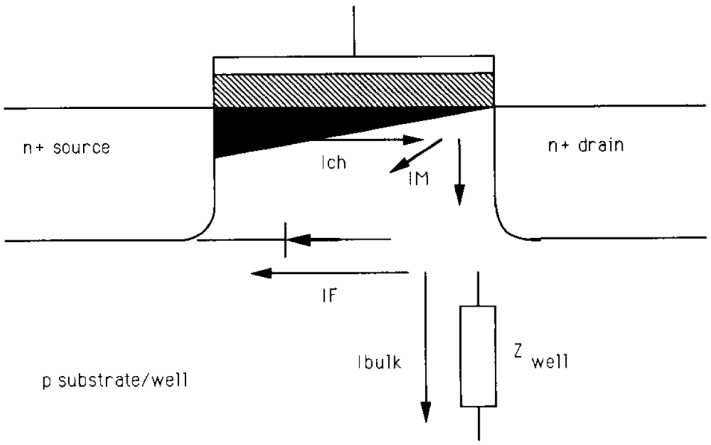

Figure 3.5: Cross-section of NMOS transistor with highlighted causes of kink [28]

in newer nodes and to increased surface scattering due to higher vertical electric fields compared to older technologies. Higher doping means lower bulk resistance while surface scattering degrades the mobility reducing impact ionization in the channel.

3.1.5. Noise

Another advantage of cooling down transistors is the reduced thermal noise. Ideally, one would expect that thermal noise decreases almost linearly with temperature. In [30], the authors reported a decrease in minimum noise figure from 1.4 dB to 0.5 dB at 30 GHz for 65-nm MOSFETs measured at 78 K while in [31] an order of magnitude decrease in the minimum achievable noise temperature, T_{min}, is observed. On the other hand, low-frequency noise, mainly dominated by flicker noise, is found to degrade at low temperatures ([32], [33]). It is generally believed that flicker noise originates from charge trapping/detrapping or from mobility fluctuations from which currently used noise models have been developed [34]. Nevertheless, this model is not valid at cryogenic temperatures and no valid explanation can be found in literature, about the reason why flicker noise increases at low temperature: [32] reports an increase in flicker-noise spectral density in NMOS while a slight decrease in PMOS (the term K decreases from 1.084 * 10^{-24} J to 0.175 * 10^{-24} J) at 77 K in a commercial 180-nm CMOS technology. [33] shows a decrease of flicker noise from room temperature to 150 K and then an increase for temperatures below 150 K. In general, flicker noise has been found to increase if the temperature is lowered, oppositely to thermal noise. From these experimental findings, high-frequency circuits would strongly benefit from cryogenic temperatures in terms of signal-to-noise ratio and noise figure.

3.2. DC Modeling of CMOS at 4 K

In order to enable complex circuit design at deep cryogenic temperatures, existing models, such as PSP [35] or MOS11 [36], must be adapted[2]. In this section, an algorithmic procedure will be shown of how to modify an existing model and shape it to work at 4 K. The adopted technology was SSMC 0.16 µm and the Spice model is MOS11 .

The starting point is a set of measurements of transistors from the above-mentioned technology. For this work, we have adopted the data presented in [37]. The model can then be adapted based on the theories discussed in the previous section.

In Fig. 3.6, 3.7, 3.8 a set of measurements at cryogenic temperature and room temperature is shown. A few things can be immediately observed, to confirm what has been discussed above:

- Increase in mobility (≈ 2 times) and threshold voltage (≈ + 0.15 V)
- Absence of kink in thin-oxide devices
- Slight decrease in series resistance at the source and drain
- Subthreshold slope increases sublinearly: for thin-oxide devices, from 87 mV/dec at RT, it reaches 22.7 mV/dec at 4 K. For thick-oxide the threshold voltage is 104 mV/dec at RT while at LHT is 35 mV/dec for low values of V_{DS} and reaches around 1 mV/dec for high values of V_{DS}. This huge discrepancy is due to the kink and the jump in current.
- In Fig. 3.7, the freeze-out effect is demonstrated by the resistance increase of an N-well resistor(the resistance was designed to be 3.5 kΩ at RT).
- Higher bulk currents for thick-oxide devices than for thin-oxide devices (Fig. 3.8), due to enhanced impact ionization.
- The I_D has a linear dependence with V_{GS}. This is a common behavior when transistors are velocity saturated.
- Drain-Induced-Barrier-Lowering is observed. In fact, for different drain potentials, the threshold voltage shifts to lower values, due to a lowering of the potential barrier at the drain side.

From these observations and findings in literature, the model can be modified as it is shown in Fig. 3.9:

1. The first plot shows the starting point, which is a simulation at T = 27 °C and standard libraries provided by the foundry. The dashed curves are the measurements at 4 K. Large mismatch is observed.
2. **Subthreshold slope**. In order to match the subthreshold slope, the temperature of the simulator was decreased down to -200°, as shown in the second plot. The simulator is now forced to extrapolate the model to such low temperature and strange anomalies appear in the curve, as highlighted in the black circle.
3. **Temperature dependences**. In order to avoid the model to wrongly extrapolate the curves, temperature dependences were switched off by zeroing the following parameters: STVFB, STPHIB, ETABETR, ETASR, ETAPH, STETAMOB, ETAR, ETASAT, STA1. This is shown in the third plot.
4. **Mobility and Threshold voltage**. To increase the mobility and the threshold voltage, the parameters BETSQR and VFBR were adapted (Fig 3.9-4th plot).
5. **Velocity saturation**. Parameter THESATR was changed to make transistors saturate at lower V_{DS} (Fig 3.9-5th plot).
6. **Surface scattering**. As explained in the previous section, surface scattering is a very important mobility degradation mechanism at low temperatures. For this reason, the parameter THESRR was varied, to enhance the effect of this degradation in simulation according to the physics. Nevertheless, the variation did not make any visible change in the curves.

[2]Existing models are certified in the military range, -55° to 125°

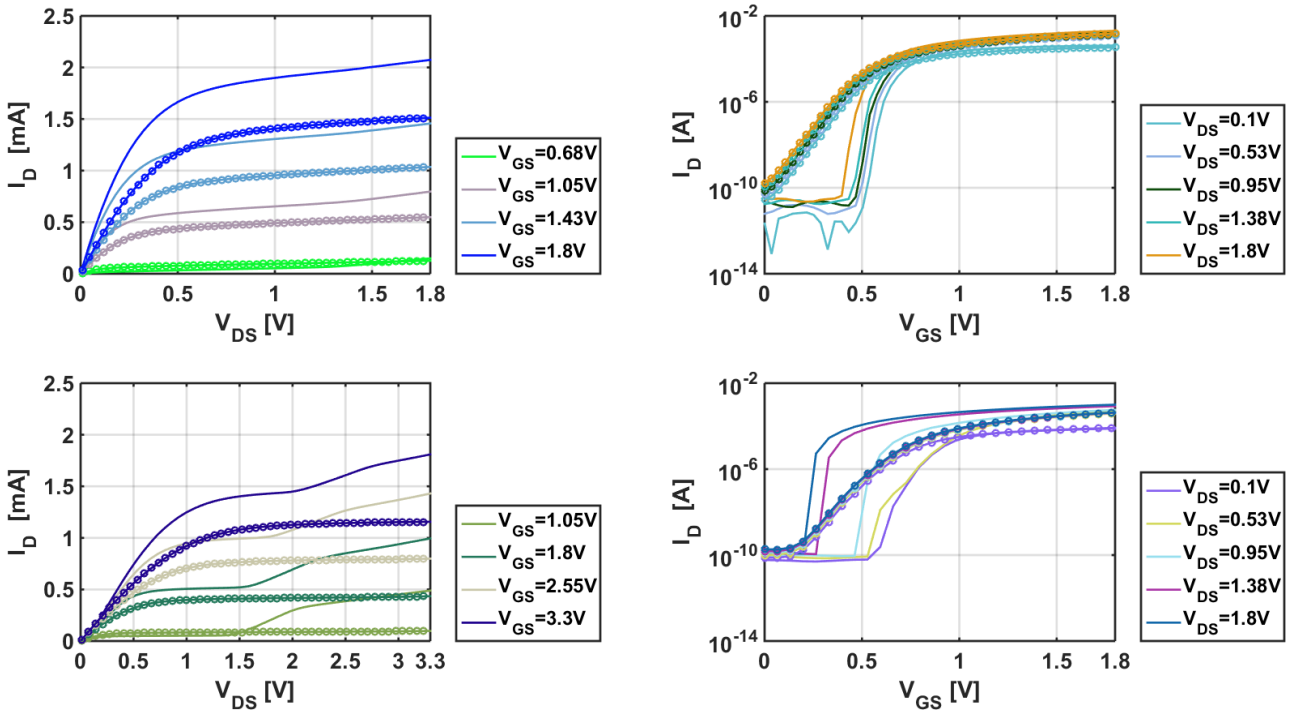

Figure 3.6: Measured curves at room temperature (circled lines) and at 4 K (thick lines) of thin-oxide (top) and thick-oxide (bottom) transistors. W_{thin}/L_{thin} = 1.6 µm/0.16 µm, W_{thick}/L_{thick} = 2 µm/0.322 µm

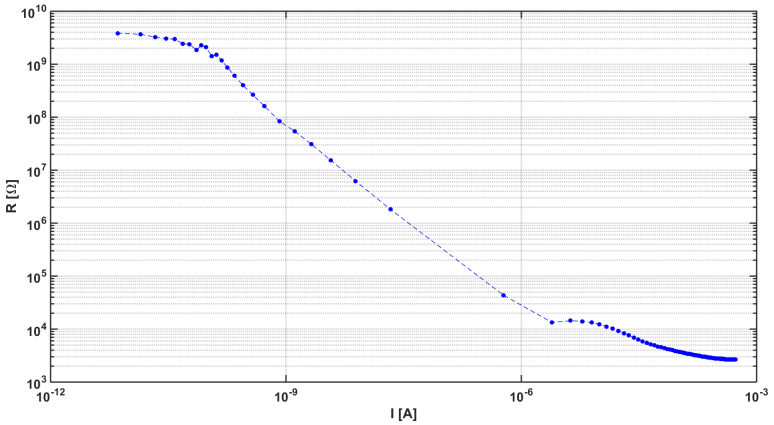

Figure 3.7: Measured N-well resistance at 4 K versus current.

7. **Impact Ionization**. In order to increase the effect of impact ionization and, consequently, enhance the bulk current parameters A1R, A2R, A3R were modified accordingly (Fig 3.9-7th plot). The consequence of this change is not directly seen in the 7th plot, because it mainly modifies the bulk current. Since thin oxide does not suffer from kink, as explained in the previous section, no effects are visible in the I_D-V_{DS} curve.

8. **DIBL**. Drain-Induced-Barrier-Lowering was augmented by changing the parameter SDIBLO (Fig 3.9-8th plot).

9. **Other parameters**, listed also in table 3.1 were modified to finely tune the fitted curves (Fig 3.9-9th plot).

10. **Kink**. For thick-oxide transistors, a non-linear resistance (100 kΩ at 1 nA) in series with the bulk contact was added. This emulated the freeze-out of carriers and therefore produced a jump in

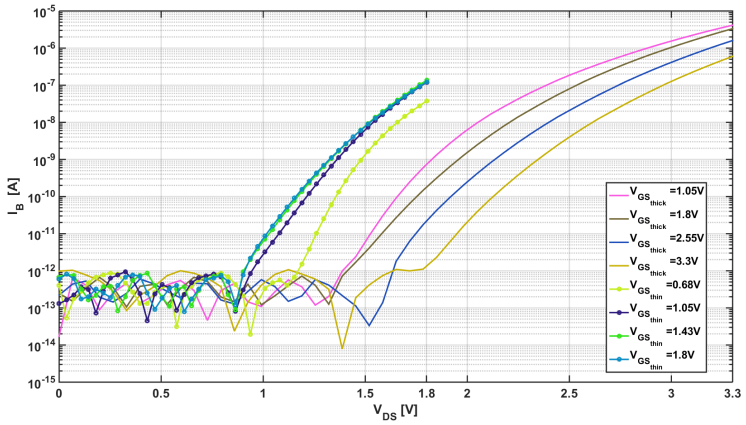

Figure 3.8: Bulk current in thick-oxide transistor, W = 2 µm, L = 0.322 µm (thick lines) and thin-oxide transistors (circled curves), W = 1.6 µm, L = 0.16 µm

Table 3.1: List of parameters modified to adapt the RT model to LHT

MOS11 parameters					
BETSQR	VFBR	THESRR	THESATR	THERR	SDIBLO
A1R	A2R	A3R	ALPR	KOR	

the curve, as measured (Fig 3.10). If added to thin-oxide transistors, no effect is observed since the bulk current is not enough to produce threshold voltage variations.

In conclusion, the same measured curves are now showed, along with simulation, in Fig. 3.10. The simulations match well the measurements. For thick-oxide transistors a mismatch is observable in the I_D-V_{DS} for low V_{GS} values. This is mainly due to the smoothing function that connects weak and strong inversion regions in MOS11. The smoothing function also needs to be changed, to perfectly match the curves at 4 K, but this would require changing the equations in the model and not only adapt the parameters, which is out of the scope of this work.

Because of the lack of high-frequency characterization, transient simulations are not reliable. Nevertheless, a good estimation of the operating point can be retrieved. However, parasitics are expected to decrease with lower temperature, therefore transient analysis with the model described above would underestimate the circuit speed: the bandwidth is expected to be larger than in simulation. DC gain, on the other hand, could be accurately estimated, as well as the operating point of the transistors.

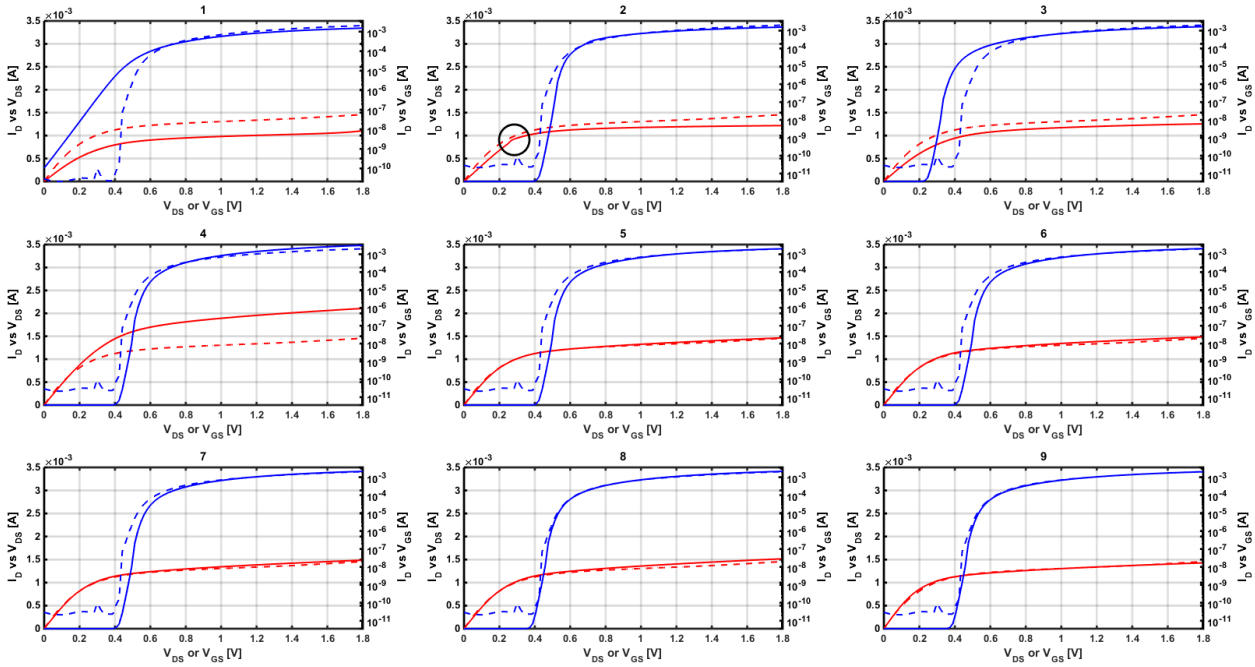

Figure 3.9: The 9 plots represents the fitting procedure step by step, along with the description given in the text. Dashed curves are measurements, thick lines are simulations. I_D-V_{GS} curves are blue while I_D-V_{DS} curves are red. The first plot compares I_D-V_{GS} and I_D-V_{DS} measured at 4 K to simulation (at 27 °C) using the model provided by the foundry, without any modifications. The second plot shows the effect of changing the temperature of the simulator from 27° C to -200° C. Third plot shows how to solve the wrong extrapolation of the simulator: zero the temperature dependences. Fourth plot shows the increase in mobility and threshold voltage. Fifth plot the degradation of the velocity saturation. Sixth plot shows the surface scattering effect. Seventh plot is taken after enhancing impact ionization. Eighth plot shows the difference after changing the parameter SDIBLO. In the ninth plot all the parameters listed in table 3.1 are changed. Dashed lines are measurements, thick lines are simulations. The transistor taken in consideration is an NMOS W/L = 1.6 µm/0.16 µm

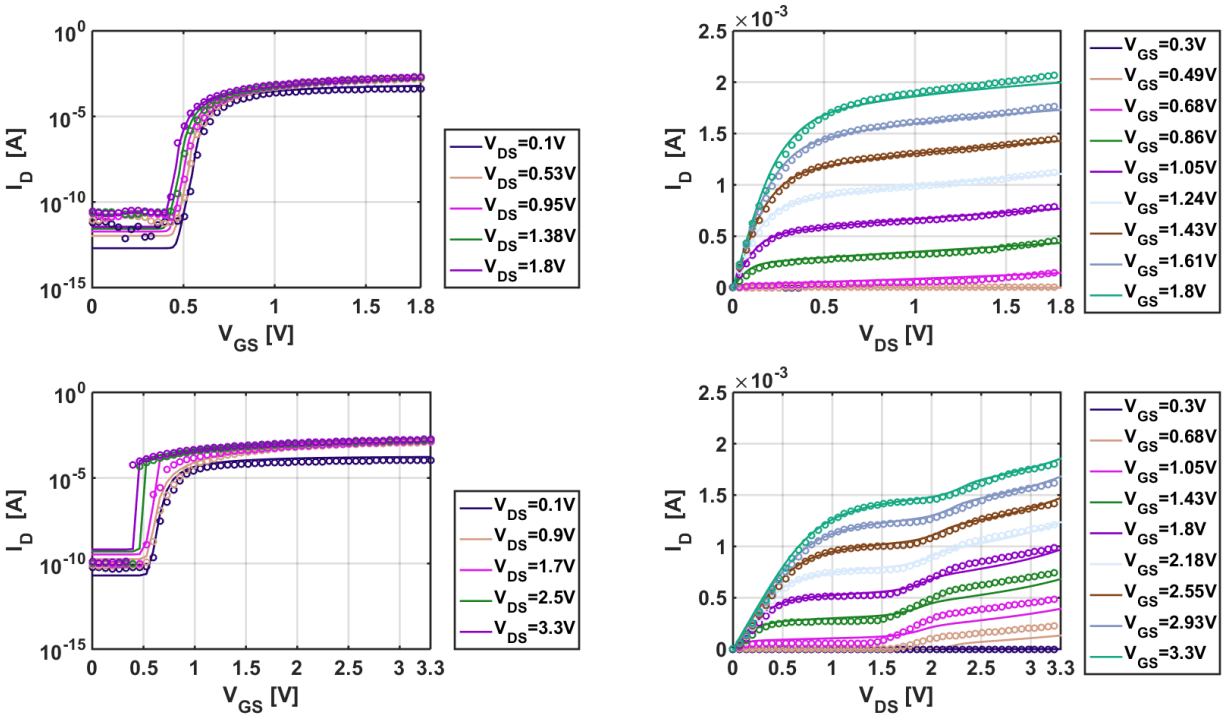

Figure 3.10: Measured curves at 4 K (circled lines) overlapped with simulated curves (thick lines) after parameter fitting. Thin-oxide size W/L = 1.6 µm/0.16 µm; Thick-oxide size W/L = 2 µm/0.322 µm

4

Design

This chapter describes the design choices to meet the specifications derived in Chapter 2.

4.1. Architecture Choice

According to the requirements derived in Chapter 2, the choice of the architecture translates into finding a circuit configuration able to match the input impedance to the line impedance and achieve a noise figure of 0.009 at the same time. This is impossible to obtain with conventional open-loop amplifier. The simplest common-source transistor, for example, provides a gate as an input which can not be used since its input impedance is much higher than 50 Ω, even if the noise can be decreased by increasing its bias current. If a shunt 50-Ω resistor is placed at the gate, input matching is achieved but noise figure would be higher than 3 dB (NF = 1 + $\frac{R_{source}}{R_{shunt}}$ + ... = 1 + 1 + ... > 2). The same would happen if a common-gate amplifier is used: although it can provide a 50-Ω input impedance, the noise figure is limited to 3 dB. In general, only with the help of active devices both parameters can be decoupled and the trade-off broken [38]. More specifically, feedback configurations or the noise-cancelling technique [39] are possible solutions. In the following sections a detailed comparison between the two alternatives is drawn, leading to the final choice. The assumption in the following comparison is that transistors are ideal, ruled by a quadratic behavior and with an infinite output impedance. This assumption enables to give a more intuitive description, without going into non-significant details.

4.1.1. Feedback configuration

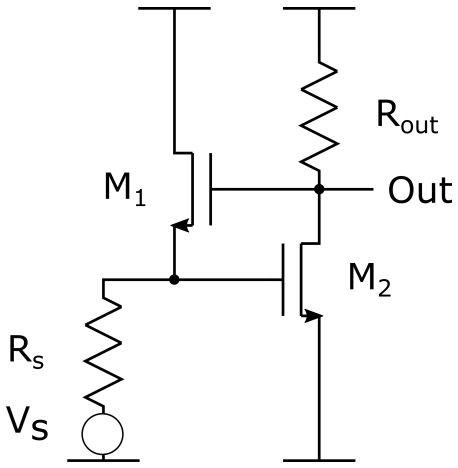

Figure 4.1: Feedback configuration

In Fig. 4.1 a possible feedback configuration is shown. This circuit achieves a noise figure below

3 dB and input impedance matching at the same time. With feedback, another degree of freedom is added in both the input impedance and noise figure expressions so that the two parameters are fully decoupled. For instance, the circuit in Fig 4.1 has an input impedance given by:

$$Z_{in} = \frac{1}{g_{m1}(1+|A|)} \tag{4.1}$$

where g_{m1} is the transconductance of M_1. Its noise figure is given by:

$$NF \approx 1 + \frac{R_{out}}{R_s}\frac{1}{|A|(1+|A|)^2}\left[\frac{1}{|A|}+\gamma\right] + \frac{\gamma}{A^2} \tag{4.2}$$

where $A = -g_{m2}R_{out}$ and γ is the excess-noise coefficient of MOSFETs. It can be noticed that the gain A enters both expressions: if a certain noise figure is required the parameters A and R_{out} can be modified; if a certain input impedance needs to be met, both g_{m1} and A can be chosen. In conclusion, there are enough degrees of freedom to achieve both requirements. Similar circuit topology exploiting the feedback can be devised. The drawbacks of these architectures are the following [40]:

- The input impedance depends on the open-loop gain of the feedback amplifier, which, on the other hand, is susceptible to process variations.
- It is difficult to ensure feedback over a wide bandwidth. Therefore, since loop gain degrades at high frequency, both impedance matching and noise figure will degrade at high frequency.

4.1.2. Noise-Cancelling technique

Fig. 4.2 depicts a simplified schematic of an example of noise-cancelling LNA.

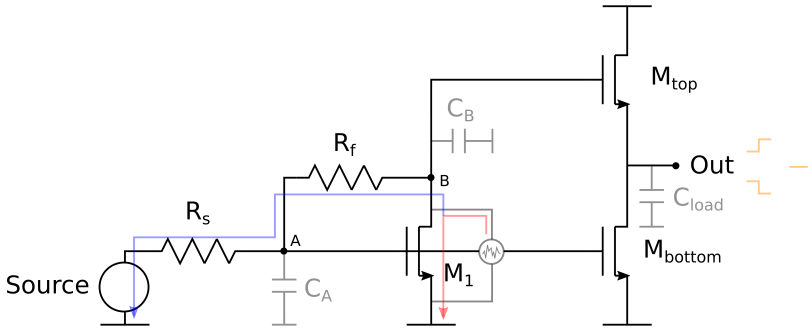

Figure 4.2: Noise Cancelling small signal circuit

This technique, first described in [39], aims at cancelling the noise of the stage providing the input matching (M_1 in Fig. 4.2, that would otherwise cause a noise figure above 3 dB. The noise is, then, dominated by the second branch (M_{top} and M_{bottom} in Fig. 4.2). The cancelling principle works as follows: the current noise of M_1, in parallel to its channel, splits in two and one part flows through M_1 (red arrow in Fig. 4.2), while the other part flows through R_f and R_s (blue arrow in Fig. 4.2). The latter produces voltage variations at node A and node B, with the same phase but different magnitude. The voltage variation at node A is amplified and inverted by M_{bottom} at node Out while the variation at node B is buffered without any inversion by M_{top} to the output. If the inverting gain from M_{bottom} to Out is tuned such that the two signals at the output have the same magnitude and opposite phase, they will cancel each other, thereby cancelling the effect of the noise of M_1 at the output (orange voltage variations in Fig. 4.2). Mathematically, the output noise due to M1 is:

$$S_{out} = S_{M_1} \cdot \left(\frac{R_f + R_s - \frac{g_{m-bottom}}{g_{m-top}}R_s}{1+g_{m1}R_s}\right)^2 \tag{4.3}$$

where $S_{I\text{-}M_1}$ $[\frac{A^2}{Hz}]$ is the power spectral density (PSD) of the current noise of M_1. By imposing that S_{out} = 0, the condition for the noise cancellation can be found:

$$\frac{g_{m-bottom}}{g_{m-top}} = 1 + \frac{R_f}{R_s} \tag{4.4}$$

The cancellation does not hold for the noise of the other devices and, consequently, the signal is degraded by the noise, but the conditions to achieve input matching and noise performance are decoupled.

The input impedance of the circuit in Fig. 4.2 is given by:

$$Z_{in} = \frac{R_f}{1 + g_{m1}R_f} \approx \frac{1}{g_{m1}} \tag{4.5}$$

where the last approximation holds for $g_{m1}R_f \gg 1$. Since the input impedance in this topology depends on a single circuit parameter (g_{m1}), achieving input matching is much easier than in the feedback counterpart for which different parameters must be kept under control, as shown in the previous section. With respect to the feedback topology, noise-cancelling offers impedance matching over a wider bandwidth and a better control over the gain. A drawback of this architecture is that the noise-cancelling condition depends on the ratio of transconductances and on the ratio of two resistances, so that a mismatch in such ratios will degrade the noise figure. However, tunability can be added to prevent the afore-mentioned issues.

For the reasons stated above, the noise-cancelling technique was adopted and specifically the architecture in Fig. 4.2 was chosen, among many other noise-cancelling architectures, because of the following reasons:

- It has a single-ended input as required in the application of this work.
- According to [40], it can be shown that it provides better noise performance than many other architectures, since it has the least number of devices needed for noise cancellation.
- Having few devices is an advantage in terms of functionality at cryogenic temperatures since the accurate behavior of cryoCMOS is not yet modelled.

In the next section, a more detailed analysis will be provided.

4.1.3. Analysis of Noise-Cancelling architecture

In order to analyse the noise-cancelling architecture in more details, the circuit in Fig. 4.2 will be taken as reference. The following assumptions are taken in the analysis:

- Noise cancellation is always achieved and Eq. 4.4 is always true.
- Input matching is met and Eq. 4.5 is always valid.
- Output impedance of transistors is neglected.

DC Gain With a simple circuit analysis it can be shown that the DC Gain of this architecture is given by:

$$\frac{V_{out}}{V_{in}} = G = \frac{1 - g_{m1}R_f - \frac{g_{m-bottom}}{g_{m-top}}}{2} = -\frac{R_f}{R_s} \tag{4.6}$$

where in the last step the noise-cancelling condition and $g_{m1} = 1/R_s$ have been used. The gain is inverting and proportional to the feedback resistor R_f.

Bandwidth The bandwidth of the circuit is determined by the capacitors at the three high-impedance nodes: A, B and Out. Capacitor C_A mainly consists of the pad capacitance and gate capacitance of M_1 and M_{bottom} in parallel, C_B includes the drain capacitance of M_1 and gate capacitance of M_{top} and C_{out}

is the parallel of the gate capacitance of a hypothetical second stage, drain capacitance of M_{bottom} and source capacitance of M_{top}. The output pole at frequency f_{out} is:

$$f_{out} = \frac{g_{m-top}}{2\pi C_{out}} = \frac{g_{m-bottom}}{(1+|G|)\cdot(2\pi C_{out})} \tag{4.7}$$

and the poles due to C_A and C_B are the roots of the following equations:

$$s^2\left[\frac{C_A C_B R_s R_f}{2}\right] + s\left[\frac{C_B(R_f+R_s)+C_A R_s}{2}\right] + 1 = 0 \Rightarrow s^2\left[\frac{C_A C_B R_s^2 |G|}{2}\right] + s\left[\frac{C_B R_s(|G|+1)+C_A R_s}{2}\right] + 1 = 0 \tag{4.8}$$

There is also a zero due to the fact that the signal has two paths (through M_{top} and through M_{bottom}) from input to output:

$$f_z = \frac{1}{2\pi\left(\frac{C_B R_s}{2}\left(1+\frac{R_f}{R_s}\right)\right)} = \frac{1}{2\pi\left(\frac{C_B R_s}{2}(1+|G|)\right)} \tag{4.9}$$

In conclusion, there is a real pole due to C_{out}, two complex-conjugated poles due to C_A and C_B and a zero. Although it can not be directly stated that the bandwidth is set by the dominant pole among the three, because of the presence of a pair of complex conjugated poles, it is still a valid assumption. In fact, the presence of the zero at a frequency very close to the conjugated poles' frequency mitigates their effect, therefore the 3-dB bandwidth can be approximated by the dominant pole of the three. The bandwidth is related to the gain of the LNA, as shown in Fig. 4.4. For this figure, the following assumptions were made:

- The pad capacitance amounts to around 600 fF
- After some iterations between Matlab and Cadence for capacitance estimation, the gate capacitance of M_1 in parallel with M_{bottom} was set to 1.1 pF. This assumption is a bit too simplistic, since this capacitance is strongly dependent on the operating point (current, bias voltage etc.) and sizing. For this reason, a first order dependence on the gain was implemented but it did not give any particular insights and therefore it was neglected for the final model.
- Capacitance C_B was assumed to be 300 fF for the same reasons stated above.

Noise The noise of the circuit is the sum of the dominating voltage noise sources:

$$S_{out} = S_{out-R_f} + S_{out-M_{top}} + S_{out-M_{bottom}} \tag{4.10}$$

where the noise of M_1 is assumed to cancel at the output. Each contribution has the following expression (all the terms are power spectral densities):

$$\begin{cases} S_{out-R_f} = S_{R_f}\cdot R_f^2 & \left[\frac{V^2}{Hz}\right] \\ S_{out-M_{top}} = \frac{S_{M_{top}}}{g_{m-top}^2} & \left[\frac{V^2}{Hz}\right] \\ S_{out-M_{bottom}} = \frac{S_{M_{bottom}}}{g_{m-top}^2} & \left[\frac{V^2}{Hz}\right] \end{cases} \tag{4.11}$$

Combining the previous equations yields to the noise figure as a function of the gain G:

$$NF = 1 + \frac{1}{|G|} + \frac{1+|G|}{|G|\cdot R_s \cdot g_{m-bottom}}\left[\frac{2}{|G|}+1\right] \tag{4.12}$$

In the last expression, it was assumed that transistors were in strong inversion and their current noise is given by $S_{current} = 4kT\gamma g_m$.

Since the noise figure is a top-level specification and it is set by the system requirements, from this equation $g_{m\text{-}bottom}$ can be retrieved as function of gain G, as shown in Fig. 4.3.

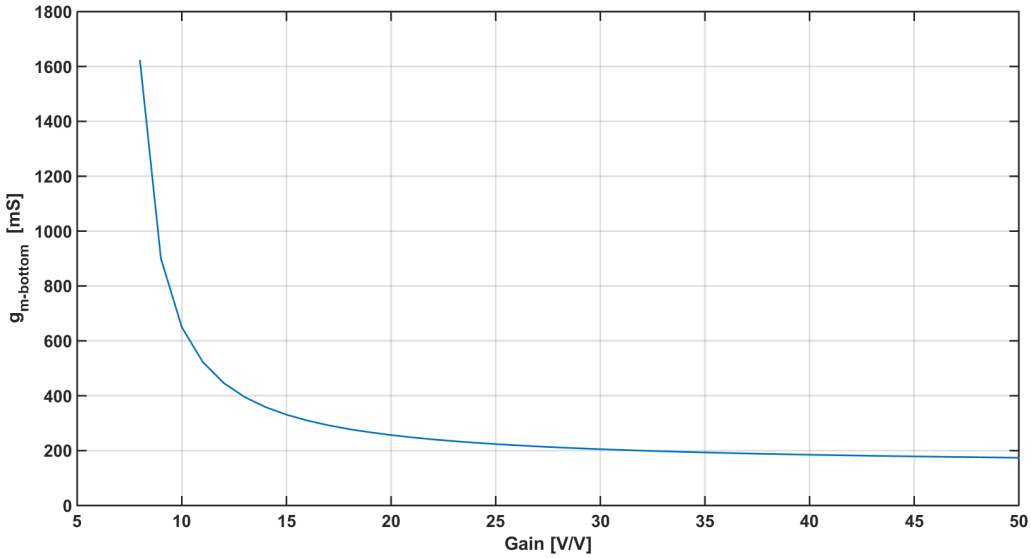

Figure 4.3: $g_{m\text{-}bottom}$ as function of gain |G|

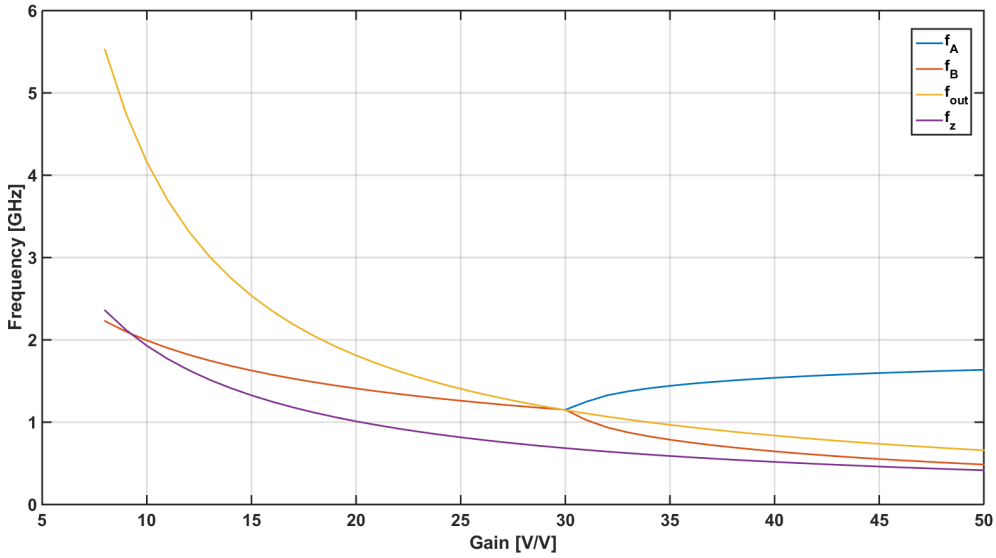

Figure 4.4: Poles and zero of LNA according to Eq.4.7, 4.8, 4.9

Table 4.1: Final top-level specifications

Noise Figure	Input noise	Input Impedance	$S11_{max}$	Bandwidth	Power/qubit
0.009 dB @ T = 4 K	40 $pV/\sqrt{Hz}$	50 Ω	-10 dB	Largest achievable	Minimum

4.2. Optimization problem: specifications for LNA

In this section, the main LNA parameters will be derived.

As mentioned in Chapter 2, the power per qubit needs to be optimized. In the next paragraph, an optimization strategy is described and, from that, the requirements of the LNA will be presented. For ease of reading, the top-level specifications are reported again, from Chapter 2.

4.2.1. Analysis

Because of the low input signal level (-135 dBm), several gain stages are used to amplify it (see Fig. 2.7). Being noise an important constraint, it is desirable that only the first stage contributes to noise and the effect of cascaded blocks, in the readout chain, is negligible. In the following analysis only the first two blocks are taken in account: the LNA and a hypothetical second stage, and the goal is to optimize the total dissipated power given a certain noise requirement. Moreover, unless otherwise stated, the temperature at which the noise is calculated is T = 4 K. Consider now Fig. 4.5:

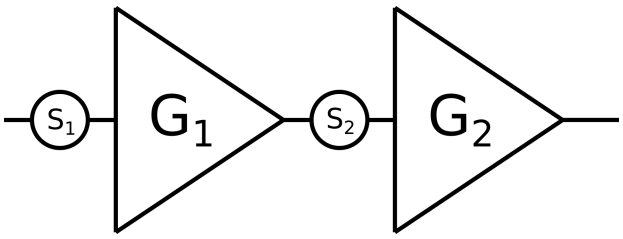

Figure 4.5: Chain of amplification

Being S_1 and S_2 the input-referred power spectral densities of the amplifiers, the input-referred noise of the cascaded system can be written as:

$$S_{tot} = S_1 + \frac{S_2}{G_1^2} \tag{4.13}$$

It can also be said that the power spent in each stage is inversely proportional to the amount of noise generated by each amplifier, assuming that both amplifiers are limited by noise. This means that:

$$P_{tot} = P_1 + P_2 + P_{constant} = \frac{k_1}{S_1} + \frac{k_2}{S_2} + P_{constant} \tag{4.14}$$

where k_1 and k_2 are design-dependent constants: if for example a ratio $\frac{g_m}{I_D} \approx 10V^{-1}$ is used, then the power can be written as $P \approx \frac{4KT\gamma \cdot V_{DD}}{10 S_{in}}$ and $k = \frac{4KT\gamma \cdot V_{DD}}{10}$. $P_{constant}$ is a constant power which is not related to noise: in case of the LNA, for instance, $P_{constant}$ is the power needed for input matching, while P_1 is the power consumed in the second branch, which is inversely proportional to noise.

In summary:

$$\begin{cases} S_{tot} = S_1 + \frac{S_2}{G_1^2} \\ P_{tot} = P_1 + P_2 + P_{constant} = \frac{k_1}{S_1} + \frac{k_2}{S_2} + P_{constant} \end{cases} \tag{4.15}$$

Combining the above two equations yields to:

$$P_{tot} = \frac{k_1}{S_1} + \frac{k_2}{(S_{tot} - S_1) \cdot G_1^2} + P_{constant} \tag{4.16}$$

where I expressed P_{tot} as a function of S_1 and G_1. P_{tot} as a function of S_1 for different values of G_1 is shown in Fig. 4.6 [1]. In the figure, k_1 and k_2 are assumed to be $2.6640*10^{-23}$ V^2W/Hz, and a g_m/I_D of 10 V^{-1} was considered. Finally, S_{tot} is $(40pV/\sqrt{Hz})^2$.

It is clear that increasing the gain of the first stage, the overall power (P_{tot}) decreases because more noise can be tolerated in the second stage and power can be saved there. In the figure, also a minimum is present for each gain. For a lower S_1, more power needs to be spent in the first stage to achieve such noise level S_1. On the other hand, if S_1 approaches S_{tot}, there is no noise budget left for the second stage, resulting in a large power to be dissipated in the latter.

At this point, it is important to notice that both power (Fig. 4.6) and bandwidth (Fig. 4.4) decrease with gain. If the gain is too low, the power of the second stage dominates and the overall power increases.

Since we are interested in P/qubit figure of merit, it is interesting to take the ratio of power and bandwidth. For a low gain, the bandwidth is very wide but the power is very large. Instead, for a high

[1] Here $P_{constant}$ is not taken in account

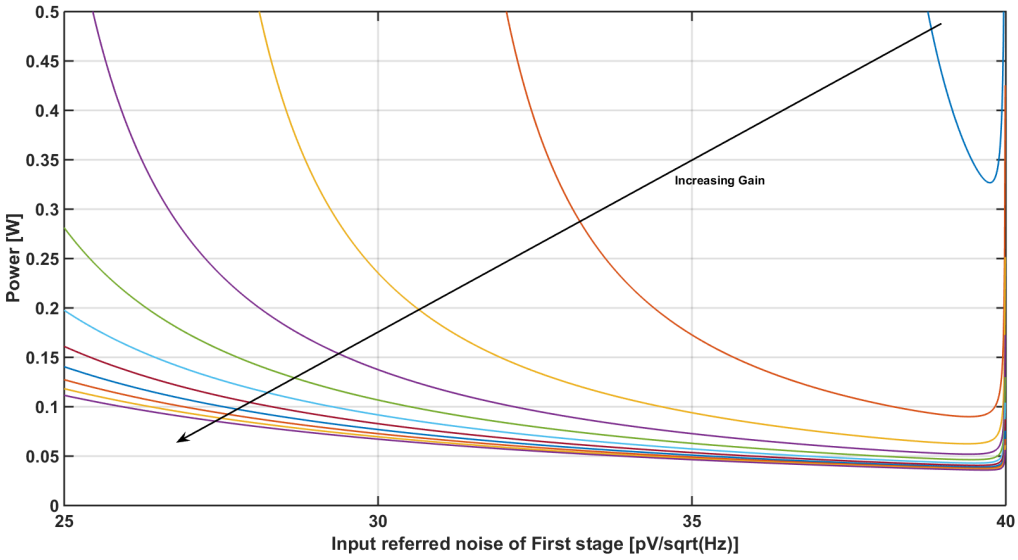

Figure 4.6: Power versus input-referred noise for different gain

gain power is lower but bandwidth is narrower. Consequently, an optimum is expected. In Fig. 4.7, the power per qubit is plotted. For this plot, all the minima in Fig. 4.6 are taken and then divided by the dominant pole in Fig. 4.4. The ratio is then multiplied by the qubit bandwidth (20 MHz), defined in Chapter 2. An optimum is found at G = 21. In table 4.2 all the main parameters found at G = 21 are summarized: $g_{m\text{-}bottom}$ is found from the noise figure (Eq. 4.12), $g_{m\text{-}top}$ is found from the noise cancelling condition (Eq. 4.4), g_{m1} from impedance matching (g_{m1} = 1/50 Ω= 20 mS). Power and bandwidth are computed from Eq. 4.6 and from Eq. 4.7 and 4.8.

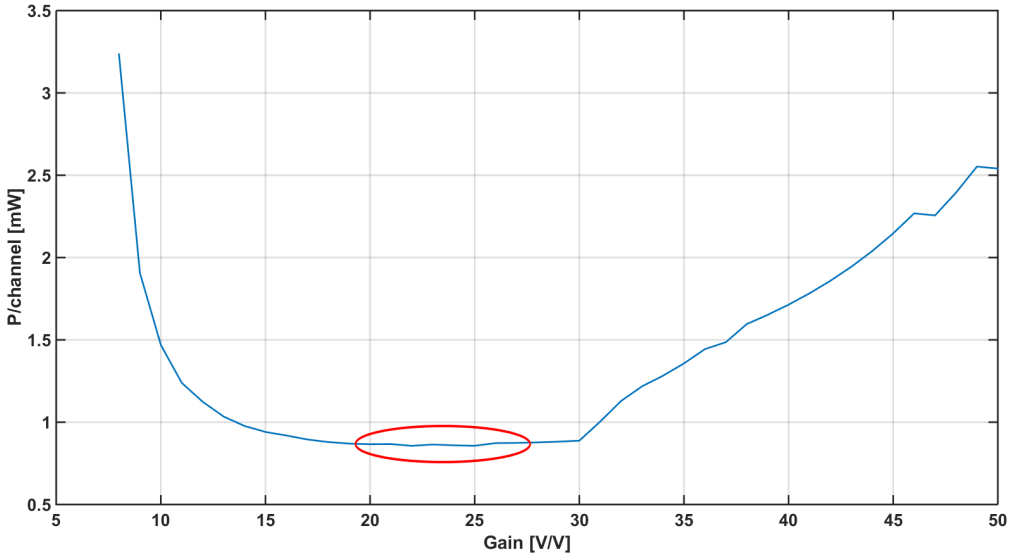

Figure 4.7: P/qubit vs Gain. In the red-circled region the optimum is found.

4.3. Circuit Implementation

After having shown that an optimum design point exists according to system level simulation, transistor-level design will be presented, according to the parameters in table 4.2. The technology for the design

Table 4.2: Final specifications for the LNA

G [V/V]	BW	NF	Input noise of LNA	Power	P/Qubit	$g_{m\text{-}bottom}$	$g_{m\text{-}top}$	g_{m1}
21	1.4 GHz	0.009 dB	$(39.4pV/\sqrt{Hz})^2$	47 mW	670 µW	250 mS	11.36 mS	20 mS

is SSMC 0.16 µm standard CMOS. Moreover, since the characterization of transistors at cryogenic temperature (described in Chapter 3) included only four sizes of transistors (namely 2.32 µm/0.16 µm, 2.32 µm/1.6 µm, 0.232 µm/0.16 µm, 0.232 µm/1.6 µm), the design was limited to the use of only those sizes or small variations of those (length changed from 0.16 µm to 0.2 µm for a few transistors).

4.3.1. LNA

Schematic In Fig. 4.8 the schematic of the designed LNA is shown:

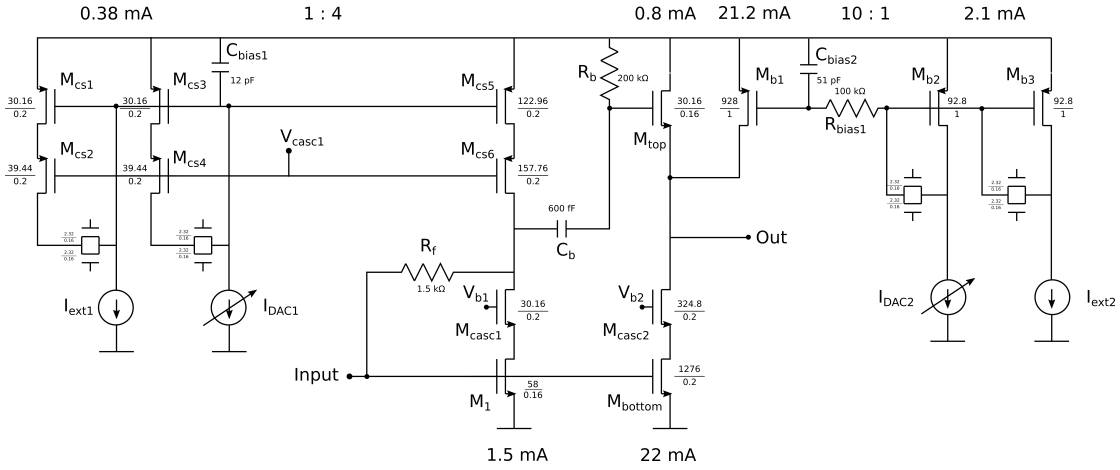

Figure 4.8: Schematic of the LNA

M_1, M_{bottom}, M_{top} are the main transistors, shown also in the small signal circuit in Fig. 4.2. They all have minimum length to preserve the bandwidth except for M_{bottom}, which has a length of 0.2 µm to increase the output impedance. This choice did not affect the bandwidth. Finally, a good trade-off between power consumption and bandwidth was found at $g_m/I_D \approx 16\ V^{-1}$ for M_{bottom} and $g_m/I_D \approx 13\ ^{-1}$ for M_1. R_f is the feedback resistor that sets the gain and input matching. The value was chosen slightly higher that what was needed to have a gain of 21, to compensate for the output impedance of the transistors M_1 - M_{casc1} and M_{cs5} - M_{cs6}. M_{b1} is a current source in parallel to M_{top} in order to tune the ratio $g_{m\text{-}top}/g_{m\text{-}bottom}$ and ensure that the noise-cancelling condition is always met. Its length was set to increase the output impedance. A cascode transistor was not used here in order to be able to increase the overdrive voltage of M_{b1} to reduce its transconductance and therefore its noise. M_{cs5}-M_{cs6} provide the current to bias the first branch, whose purpose is impedance matching. Cascode transistors are added to enhance the output impedance of M_1, M_{bottom}, M_{cs5}, M_{cs2} and M_{cs4}. C_b and R_b form a high-pass filter (pole at 1.4 MHz) to AC-couple M_{top} to the first branch. C_{bias2} and R_{bias1} are used to decrease the impedance at the gate of M_{b1} and decouple the effect of the transconductance of M_{b2} together with the gate-to-drain capacitance of M_{b1} ($C_{gd\text{-}b1}$). In fact, if C_{bias2} and R_{bias1} were not added, $C_{gd\text{-}b1}$ would create a low impedance path between output and the diode-connected M_{b2} or M_{b3}, affecting the transfer function. The currents through M_1 and M_{b1} are made tunable both externally or with IDAC (described in the next paragraph).

Layout Care was taken while layouting the LNA to keep parasitics to the minimum. In particular, capacitors would degrade the bandwidth and resistors in series with the signal would degrade the noise figure. To take care of the former the following precautions have been adopted:

- The signal has been routed in the top metal (which is Metal 5 in this technology). The ground and supply lines have been routed in the lowest metals (Metal 1 and 2) to decrease the parasitic capacitances to the signal.

- An ad-hoc PAD has been designed for low capacitance because of the lack of RF pads in the technology.

To take care of the parasitic resistances:

- The input node was placed immediately next to the input pad.
- The design was made as compact as possible in order to avoid long traces.
- Many vias were placed in parallel to reduce their series resistance.

The layout of the LNA is shown in Fig. 4.9.

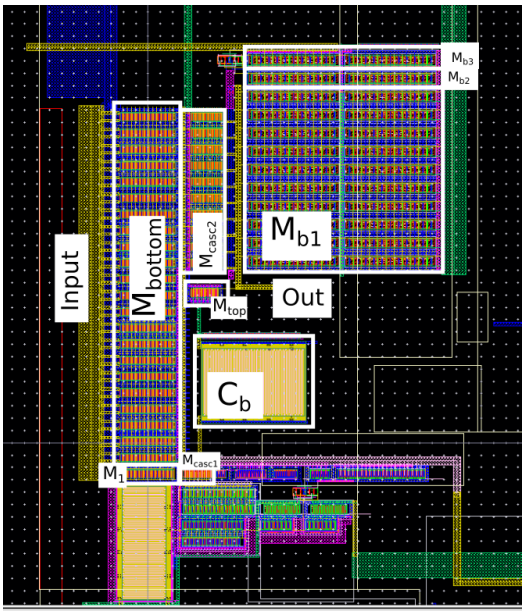

Figure 4.9: Layout of the LNA

Simulations The schematic simulations were run with a preliminary cryogenic model that was simpler than the one explained in Chapter 3, because of time constraints. In particular the temperature of the simulator was kept at 27 °C, losing the subthreshold fitting. Since the noise figure depends on the operating temperature, at 4 K the noise figure would be 0.009 dB while at room temperature (RT) the noise figure is 0.6 dB. Since it is not possible to run noise simulations at LHT, 0.6 dB is the correct value to aim at during simulations. The currents are tuned until the noise cancellation is achieved and the final values are shown in Fig. 4.8. In Fig. 4.10 the schematic and post-layout simulation of the LNA's AC response is shown. A voltage gain of 21.7 and a 3-dB bandwidth of 1.13 GHz are achieved in schematic, while a gain of 19 and a 3-dB bandwidth of 830 MHz is observed after layout. Schematic and layout were simulated with the current setting that provides the minimum noise figure, shown as a function of frequency in Fig. 4.11. The DC operating point in the two cases are slightly different because of extraction of parasitic resistances along with the capacitances. For this reason, the gain is slightly lower as well as the bandwidth. The 40 dB/dec roll-off is due to the output pole and one of the poles due to nodes A and B. The high-pass filter is given by the AC coupling at the input of the amplifier and by C_b-R_b in Fig. 4.8.

In Fig. 4.11 the noise figure over frequency is shown. The dashed line defines the limit for the NF to be 0.6 dB at RT. As it can be seen, this low level of NF can be hold up to 400 MHz and up to 1 GHz with a degradation of 0.4 dB. For initial experiments, because of the lack of a large number of qubits, this behavior over frequency is thought to be enough.

In conclusion, table 4.3 shows the performance of the LNA after transistor design. The numbers match pretty well with table 4.2, demonstrating that the specifications are met at nominal corner. A few differences can be spotted:

- The bandwidth is slightly lower due to the simplistic estimation of the parasitic capacitance
- $g_{m\text{-bottom}}$ is almost 50 % larger than expected to compensate for the finite output impedance of transistors, that was not taken in account initially.

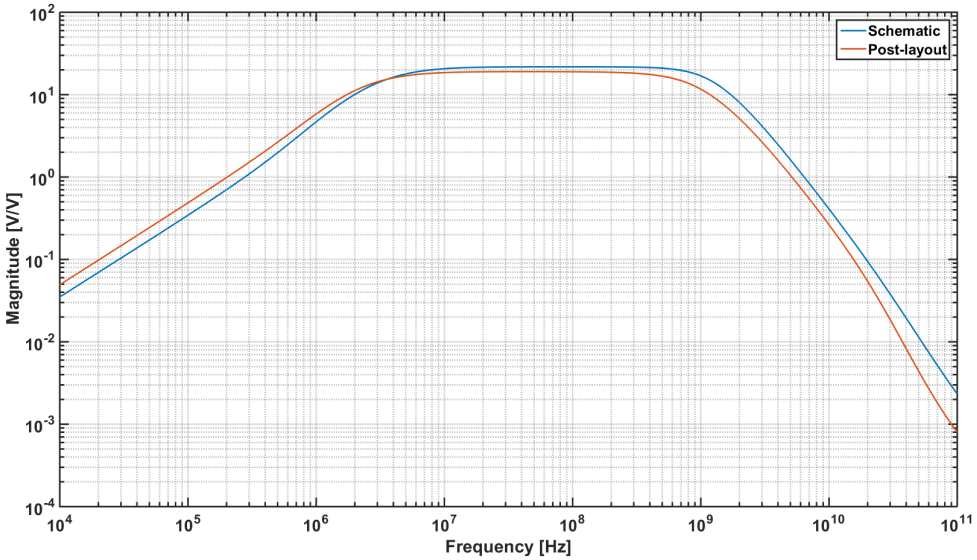

Figure 4.10: Schematic and Post-Layout AC simulation of the LNA at nominal corner

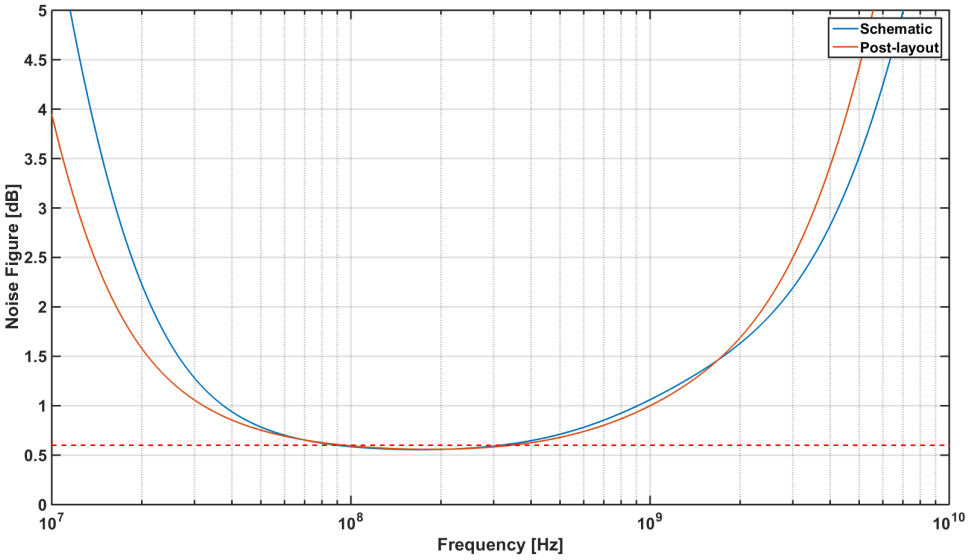

Figure 4.11: Schematic and Post-Layout simulation of NF over frequency at nominal corner. The dashed line represents the target.

Table 4.3: LNA performance after transistor-level design. First line corresponds to schematic design while the second to layout.

Gain [V/V]	Bandwidth	NF	Power	P/Qubit	$g_{m\text{-bottom}}$	$g_{m\text{-top}}$	g_{m1}
21.7	1.13 GHz	0.55 dB	49.6 mW	1.06 mW	348.8 mS	11.6 mS	22.8 mS
19.2	830 MHz	0.55 dB	50 mW	1.7 mW	/	/	/

4.3.2. Bias Circuit

Schematic A self-biased current generator [23] was implemented to produce all the bias voltages needed in the circuit. The reference current is 50 µA and is inversely proportional to the resistor R_b. A path to an external current generator is also provided as a safety option. The schematic is shown in Fig. 4.12.

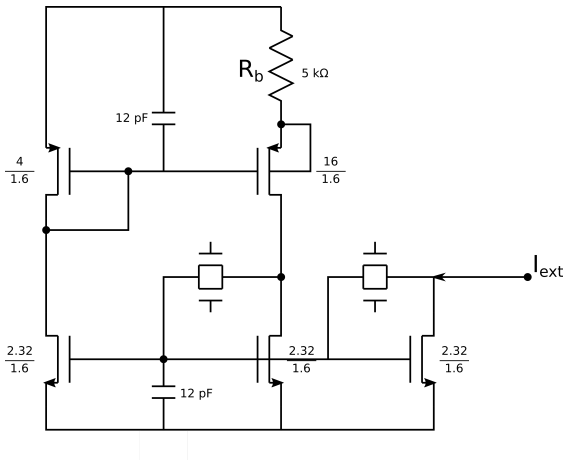

Figure 4.12: Schematic of Bias circuit

The current spread of the generator is shown in Fig. 4.13 and shows a variation of +28 µA at 'fast' corner and -14 µA at slow corner.

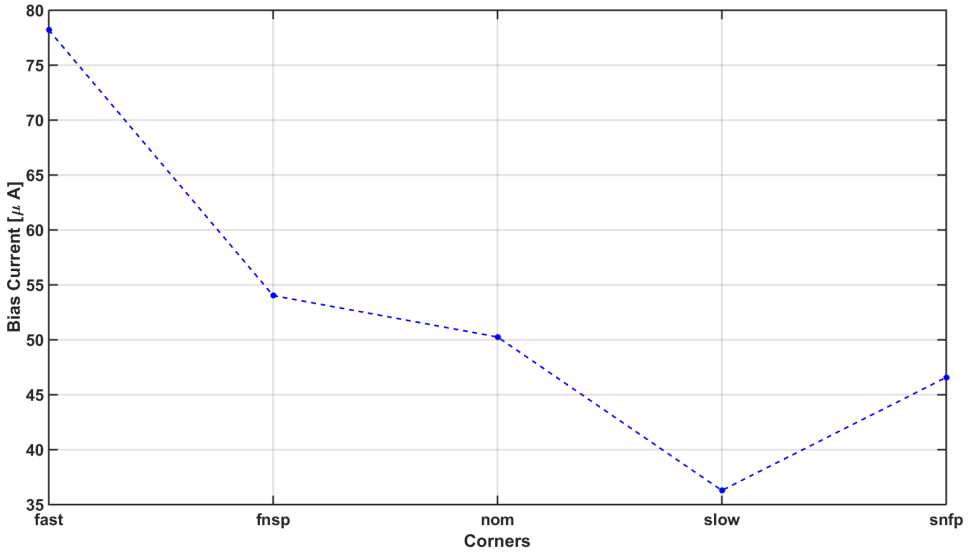

Figure 4.13: Bias current spread over corners. Schematic simulation.

In the next sections, it will be shown that this spread does not stop the circuit from achieving the noise-cancellation condition.

Layout The layout of this block does not present any special features.

4.3.3. IDAC

Because of the many unknowns at LHT, many tuning knobs were implemented in order to be able to bring the circuit back to its correct operating point, ensuring noise cancellation after process spread and tune the input matching to the optimum point. Among these, two IDACS were designed to tune

the current of M_{b1} and M_{cs5}: the first to always meet the noise-cancelling condition, regardless of mismatch and spread, the latter to ensure input impedance matching. In particular, the first IDAC will tune I_{DAC2} and the other I_{DAC1} in Fig. 4.8.

Schematic The schematic is shown in Fig. 4.14. In order to determine the range of the DAC for

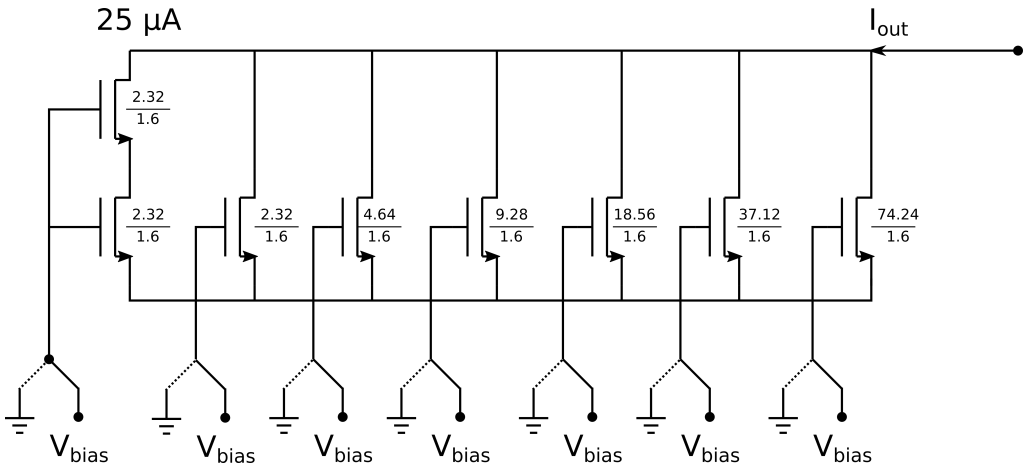

Figure 4.14: IDAC schematic

noise cancelling, in Fig. 4.15 the noise figure versus I_{ext2} is shown, including the spread of the bias current and of the LNA, in all corners. In order to recover the NF in all scenarios, a range close to 3 mA was chosen. Fig. 4.16 shows the noise figure versus I_{ext2} at nominal corner: the noise figure when the noise cancelling condition is met is 0.55 dB (little margin was taken in the design). Recalling that the upper limit is 0.6 dB, an error of 0.05 dB is tolerable. From the figure, it can be seen that this translates in a range of $\pm$ 25 µA from the minimum, setting the maximum resolution to 50 µA. A resolution of 25 µA was chosen to make sure unknown effects at cryogenic temperature were included. In conclusion, the DAC has 7 bits, 25 µA resolution and 3.2 mA full-scale range. To save time in the design, the second DAC is identical to the one just described. In Fig. 4.17 a Montecarlo simulation was run (300 runs): after design and including mismatch, the final LSB (calculated as (full range)/2^7) is 24.1 µA and the maximum variation is within $\pm$ 3 % of the mean value. The DAC was tested along with the bias circuit and connected to the LNA, therefore all the loading effects were taken in account.

Layout Dummies have been added on both sides of the transistor row to prevent too large mismatch errors.

Simulations In Fig. 4.18, the noise figure is plotted for different bit configurations of I_{DAC2}, at different corners (schematic simulation). The dashed red line shows the upper bound for the NF, which is 0.6 dB at room temperature. It can be seen that, over process variations, the IDAC is effective and the noise cancelling condition can be met in each scenario.

The same can be said for the reflection parameter S11. In Fig. 4.19 S11 versus bit configuration over corners is shown (schematic simulation) while varying the current I_{DAC1}. It is clear that it can always be tuned back to its minimum value. Although the effectiveness of the IDAC was proven at a frequency spot of 200 MHz, the behavior over frequency does not change by using the DAC from Fig. 4.11.

4.3.4. Serial-to-parallel Shift Register

In Fig. 4.20 the top-level schematic of the shift register is shown. Its purpose is to allow external programming of the different settings for the chip. In particular, it can be programmed to select between external or internal bias current in the bias circuit (Fig. 4.12) or in the LNA. The total number of bits is 17, 7 bits for each of the two DACs and 3 bits to select each current source. The bit configuration is shown in Tab. 4.4. It features two rows of flip-flops: the first row receives serially the bits from an external FPGA and, when everything is received, the FPGA activates the ”Load” signal to load the

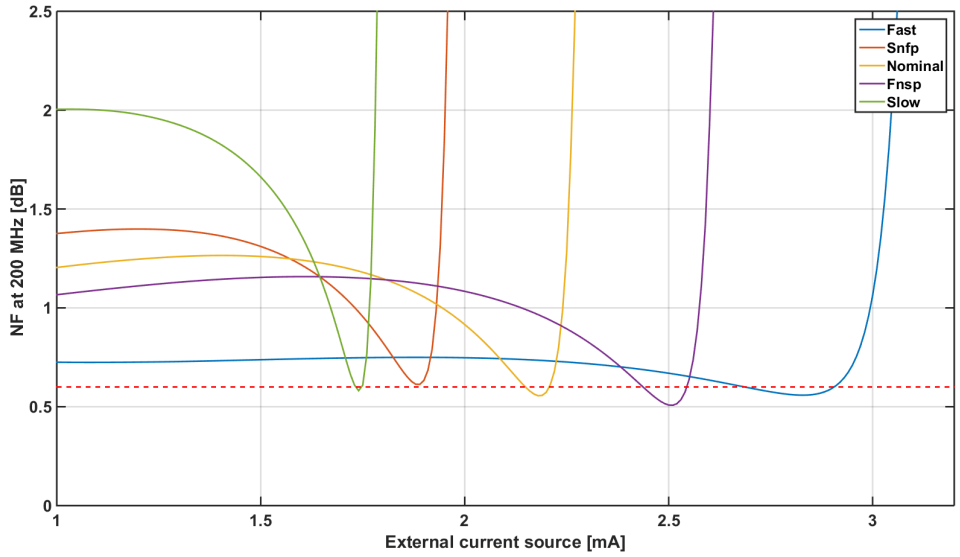

Figure 4.15: Schematic simulation of NF versus I_{ext2} at different corners. Bias and LNA spread are included. The NF was taken at a frequency of 200 MHz, which is the start of the valid bandwidth and the frequency at which the existing qubit is placed, as explained in 2.3.6.

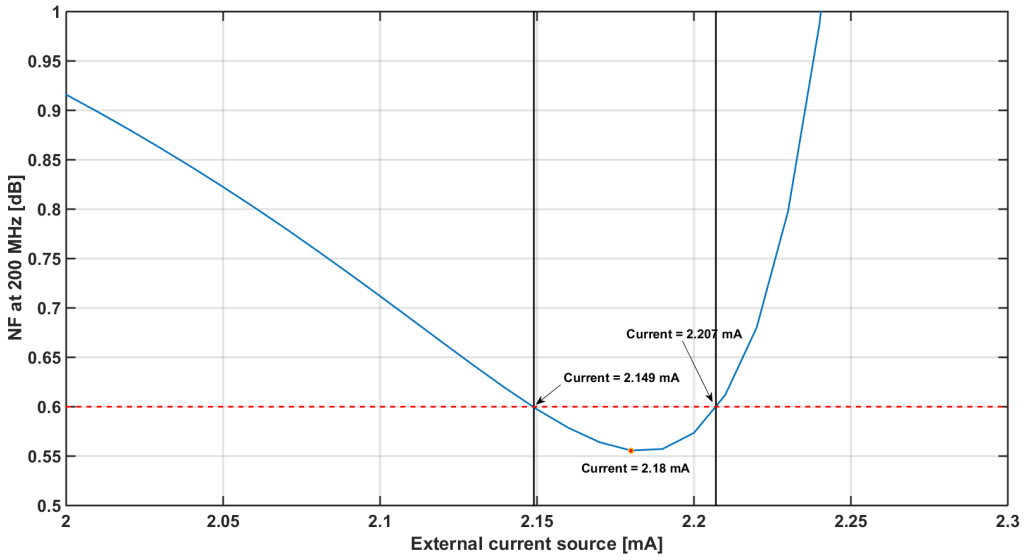

Figure 4.16: Schematic simulation of NF versus I_{ext2} at nominal corner. The NF was taken at a frequency of 200 MHz, which is the start of the valid bandwidth and the frequency at which the existing qubit is placed, as explained in 2.3.6.

programmed word into the upper row. An array of multiplexers is connected at the output of each flip-flop in the upper row to implement a ”default” state: in case the shift register did not work at LHT, a hard-wired default can be sent to the multiplexer by asserting the signal ”Sel”: this will make sure that the output of the shift-register is set to all 0s and the flip-flops are neglected. This will enable full testability because external current sources would replace the DACs.

Table 4.4: Bits placement in the shift register

Bias current	LNA current left	LNA current right	DAC1	DAC2
1 bit	1 bit	1 bit	7 bits	7 bits

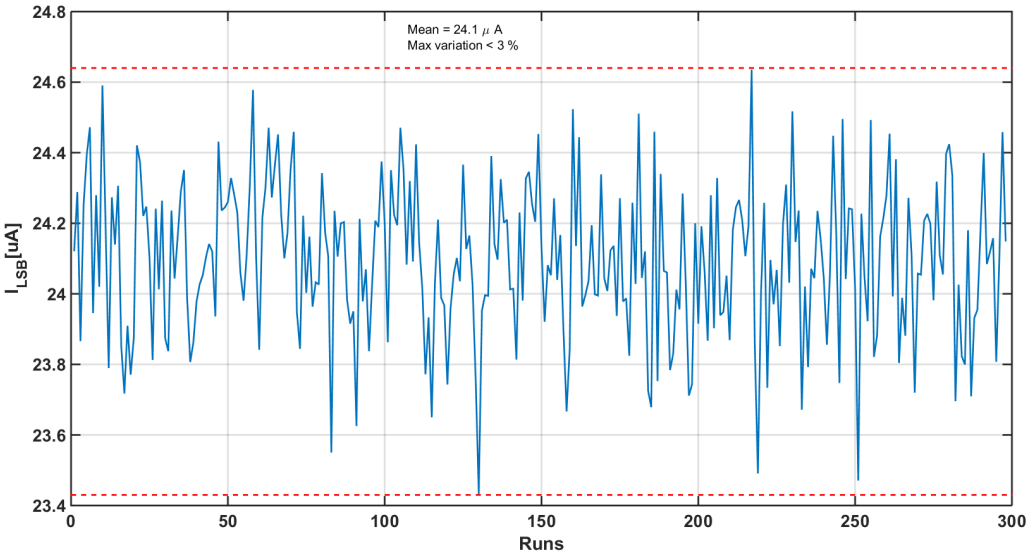

Figure 4.17: Montecarlo simulation of DAC schematic showing the variation of the LSB after mismatch of transistors.

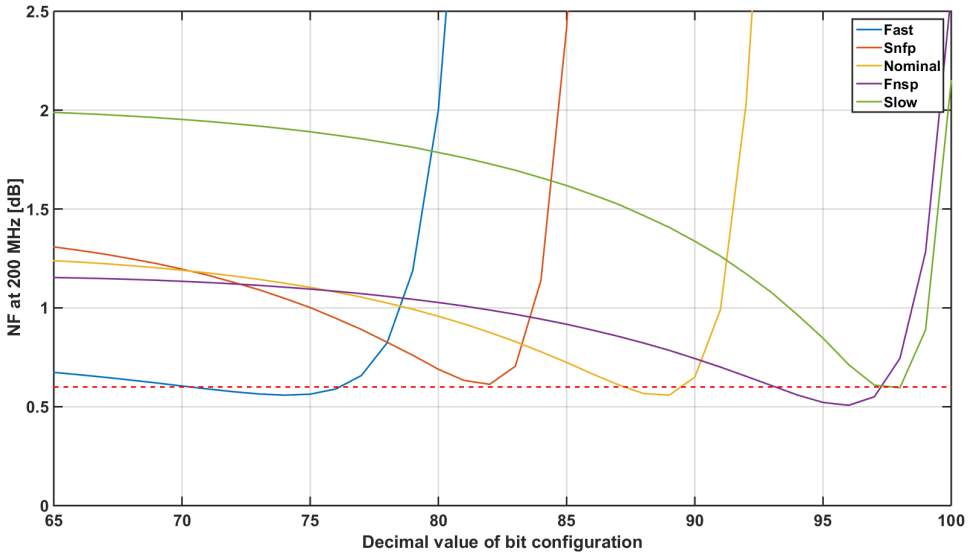

Figure 4.18: Schematic simulation of noise Figure vs bit configuration at different corners. The NF was taken at a frequency of 200 MHz, which is the start of the valid bandwidth and the frequency at which the existing qubit is placed, as explained in 2.3.6.

4.3.5. Additional Gain Stages

Since the signal to be amplified is extremely small (-135 dBm), additional gain stages were added. These stages will be useful also for noise measurement purposes, since the output-referred noise of the LNA is around $840\frac{pV}{sqrt(Hz)}$ and detecting it, with low gain is extremely difficult[2].

Schematic The schematic is shown in Fig. 4.21: a simple common source transistor with a diode as load was chosen. This architecture provides single-ended input and a wide bandwidth, that were the main requirements. M_{bottom} is the input transistor and M_{top} is the load. The input is AC coupled and the bias is set by the resistor R_{bias}. M_{CS} is a current source in parallel with M_{top} in order to better control the gain (see eq. 4.17). All the other transistors are for bias purposes. With a simple small-signal

[2]considering the noise floor of the state-of-the art benchtop instrumentation

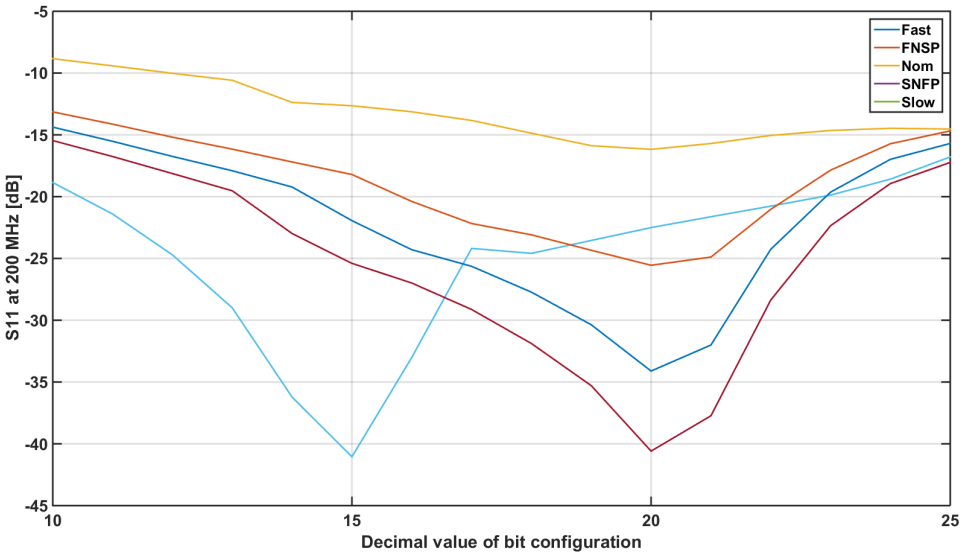

Figure 4.19: Schematic simulation of S11 vs bit configuration at different corners. S11 was taken at a frequency of 200 MHz, which is the start of the valid bandwidth and the frequency at which the existing qubit is placed, as explained in 2.3.6.

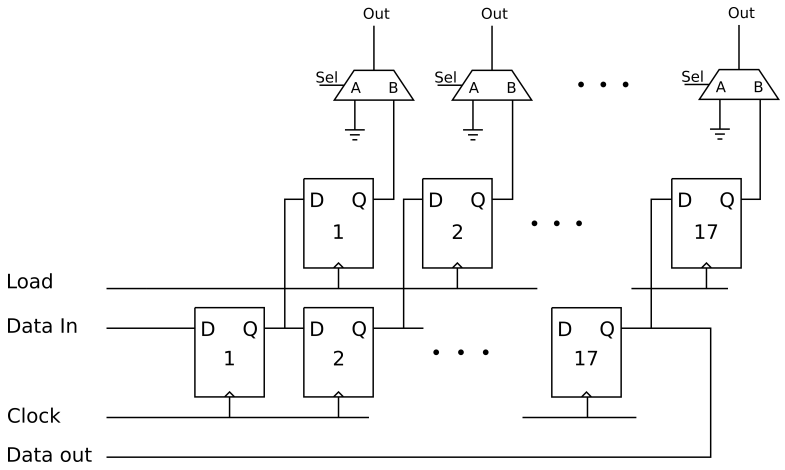

Figure 4.20: Schematic of Shift Register

analysis the gain can be found (neglecting finite output impedances):

$$G = -\frac{g_{m-bottom}}{g_{m-top}} \tag{4.17}$$

In order to set the gain the following assumptions were made:

- From previous measurements (made by B. Patra) with the same setup (described in Chapter 5), a loss of -15 dB at 6 GHz was reported for the long signal cables from LHT to RT. In the worst case, the same attenuation was assumed here, and 15 dB of gain is needed, at least, to recover back this loss.
- The overall output noise of the LNA is around $840\frac{pV}{sqrt(Hz)}$ at LHT. Noise measurements were supposed to be performed with a spectrum analyzer whose minimum noise floor is -155 dBm/Hz. At the output of the LNA there will be the noise of the 50-Ω source resistance amplified plus the noise of the LNA. The sum corresponds to around -158 dBm/Hz[3] which is smaller than the limit

[3] $S_{in} = (40pV/\sqrt{Hz})^2 + 4kT_{4K}R_s = 1.2640 \cdot 10^{-20}V^2/Hz$. In dBm/Hz, $S_{dBm/Hz} = 10 * log_{10}(1.2640 \cdot 10^{-20}/50\Omega/1mW) = -186dBm/Hz$. At the output, considering the gain of the LNA, 28 dB, $-186dBm/Hz + 28dB = -158dBm/Hz$.

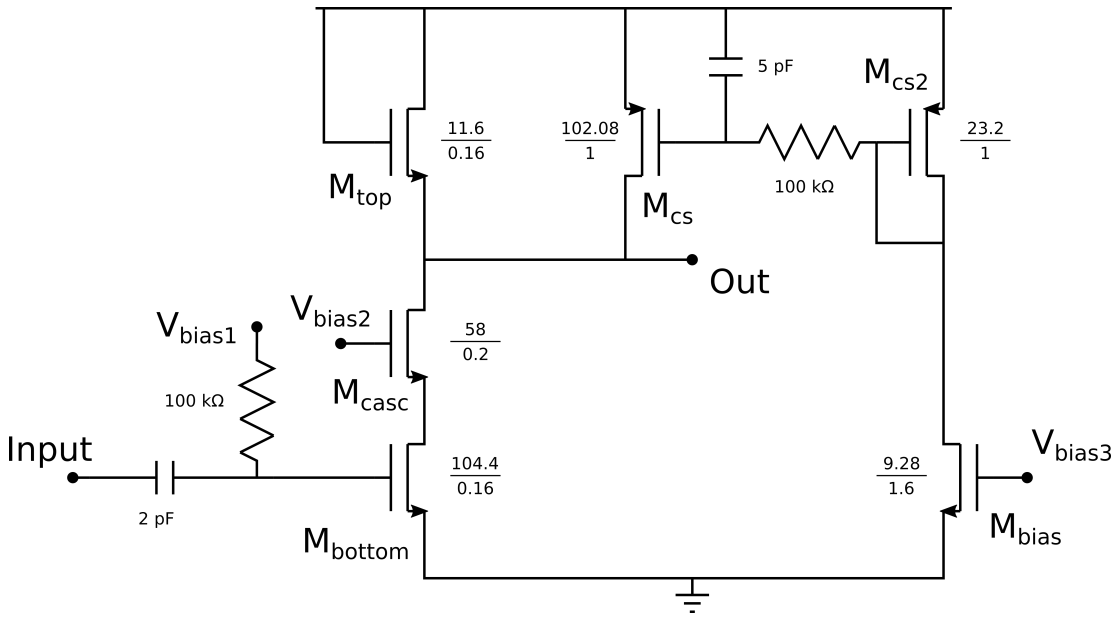

Figure 4.21: Gain stage schematic

of the analyzer. Therefore some gain is needed to be capable of measuring the noise.

- Some gain will be lost into the driver of the 50-Ω line as explained in the next section. Also this 'loss' was taken in account.

For these reasons, an overall gain of 30 dB was chosen, which translates in a gain [V/V] close to 30. Therefore 3 stages were added in series and a gain slightly larger than -3 for each was chosen.

The bandwidth is defined by the capacitor at the output node and is:

$$f_p = \frac{g_{m-top}}{2\pi C_{load}} \tag{4.18}$$

It was set to 3 GHz in order to preserve the overall bandwidth to be around 1 GHz. Assuming a C_{load} of 300 fF, then $g_{m\text{-}top}$ must be around 6 mS. Considering that the gain per each stage must be around -3, then $g_{m\text{-}bottom}$ should be around 18 mS. The input-referred noise of this architecture is given by:

$$S_{in} = \frac{4kT}{g_{m-bottom}} \cdot \frac{1+|G|}{|G|} \tag{4.19}$$

where G is defined in eq. 4.17. In order for the noise contribution to be negligible, the input-referred noise of the second stage should be around $1\frac{nV}{\sqrt{(Hz)}}$[4]. This translates into a minimum value for $g_{m\text{-}bottom}$ of 220 µS much smaller than that of the bandwidth's requirement. Consequently, the current required by those stages is not determined by noise requirements, but rather by bandwidth requirements.

Simulations In Fig. 4.22 the schematic and post-layout AC simulation of the circuit are plotted. It shows an achieved gain of 3.2 and a bandwidth of 3.2 GHz for schematic and a a slightly lower bandwidth (2.6 GHz) after layout, due to parasitics.

From 'noise' simulation, the overall input-referred noise amounts to 1.5 $nV/\sqrt{(Hz)}$ at 200 MHz at RT (after layout), which corresponds to 170 $pV/\sqrt{(Hz)}$ at LHT. As expected, the additional stages do not affect the noise of the full chain. In Fig. 4.23 the power spectral density (PSD) is shown for both schematic and layout, confirming the previous statement. The power consumption of a single stage is close to 5 mW which corresponds to around 10 % of the LNA's consumption.

[4]This number come from the optimization algorithm where the noise has been allocated in the two stages to optimize for power.

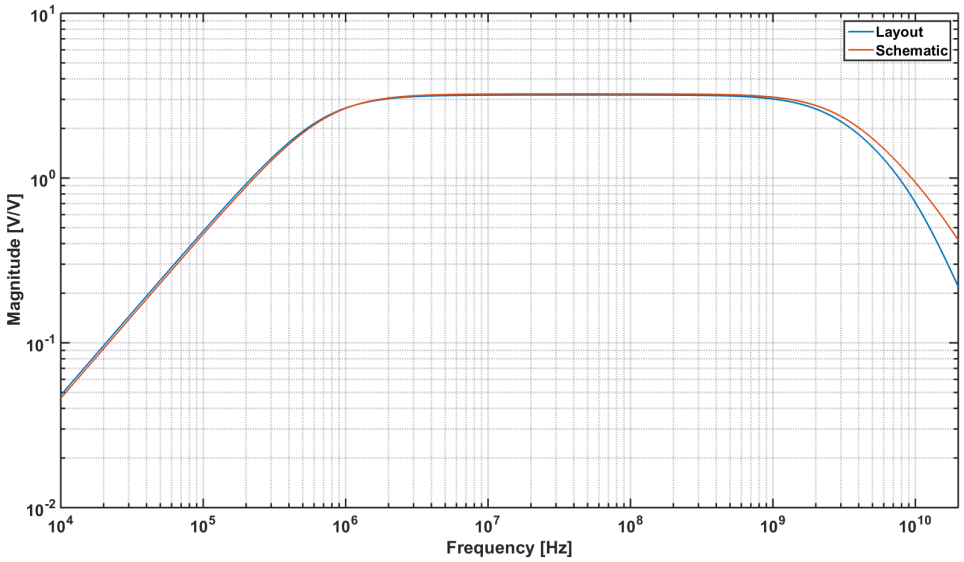

Figure 4.22: Schematic and Post-Layout simulation of the Gain versus frequency.

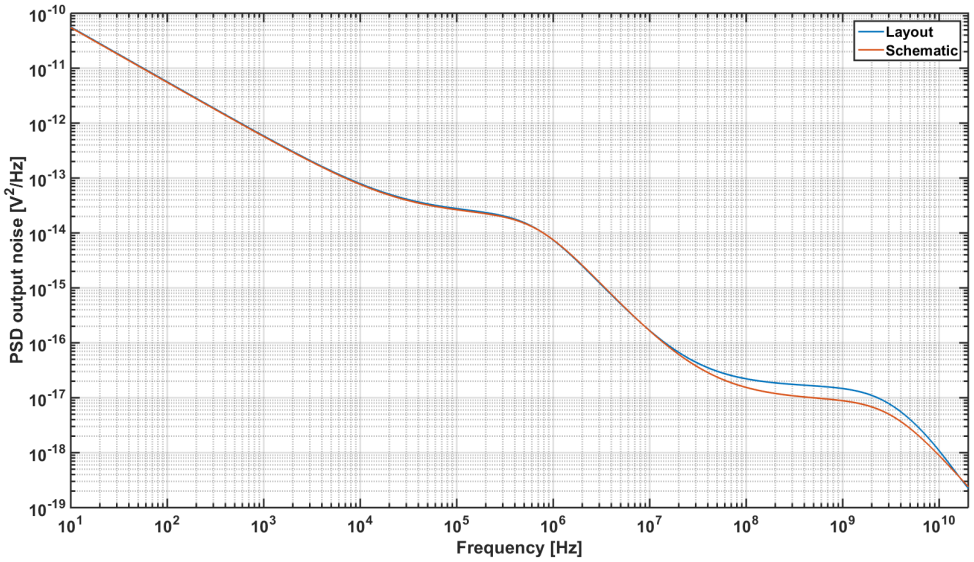

Figure 4.23: PSD of second stage amplifier. Schematic and Post-Layout simulation.

Layout The layout of this stage is not as critical as the LNA but special care has been taken to make the signal path as short as possible and to avoid parasitic capacitances.

4.3.6. Driver 50 Ω

In order to drive the long coaxial cables with a 50-Ω characteristic impedance, an output driver has been added at the end of the chain.

Schematic In Fig. 4.24 the schematic of the driver is shown. The common drain transistor acts as a buffer and provides an output impedance equals to $1/g_m$ and, depending on the bias current (set externally), it can provide 50-Ω output impedance, thus achieving output matching to the coaxial cable connected to the output.

In Fig. 4.25, s22 is shown. According to the post-layout simulation, the driver can provide output

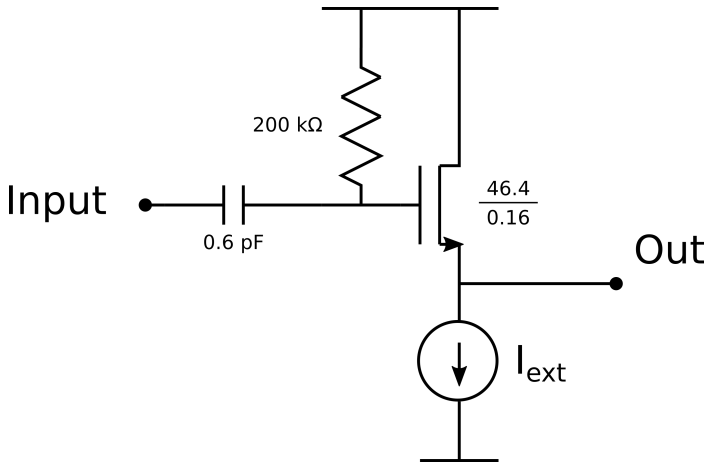

Figure 4.24: Driver Schematic

matching of -10 dB over a 3 GHz bandwidth. In this simulation a pad capacitance of 1 pF was included.

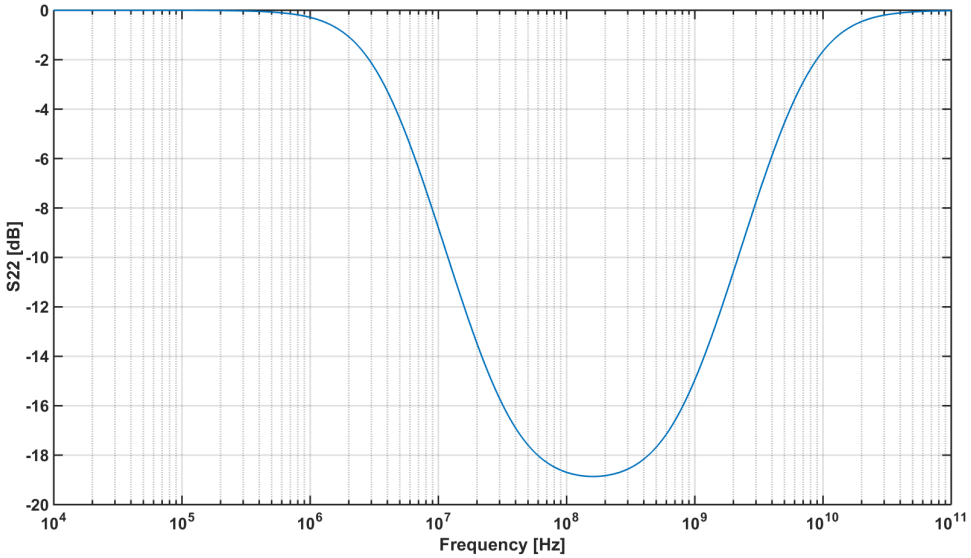

Figure 4.25: Layout simulation of s22.

Layout The layout of this block is not critical and no special features need to be commented.

4.3.7. Full chain

Full Schematic The top level of the full schematic is shown in Fig. 4.26. It features the LNA, three gain stages and an output driver for output matching. Bias circuit provides the voltages (red lines) while the shift register provides 7 bits to each DAC (blue arrows) plus 3 selection bits (that are not shown to make the schematic readable), to choose between external current source and DAC or between external and internal current source for the bias circuit.

Layout The layout of the complete chip is shown in Fig. 4.27: As it is shown, all blocks carrying the signal are very close to each other. The LNA was placed as close as possible to the input pad, in order to avoid parasitic resistance in series with the signal.

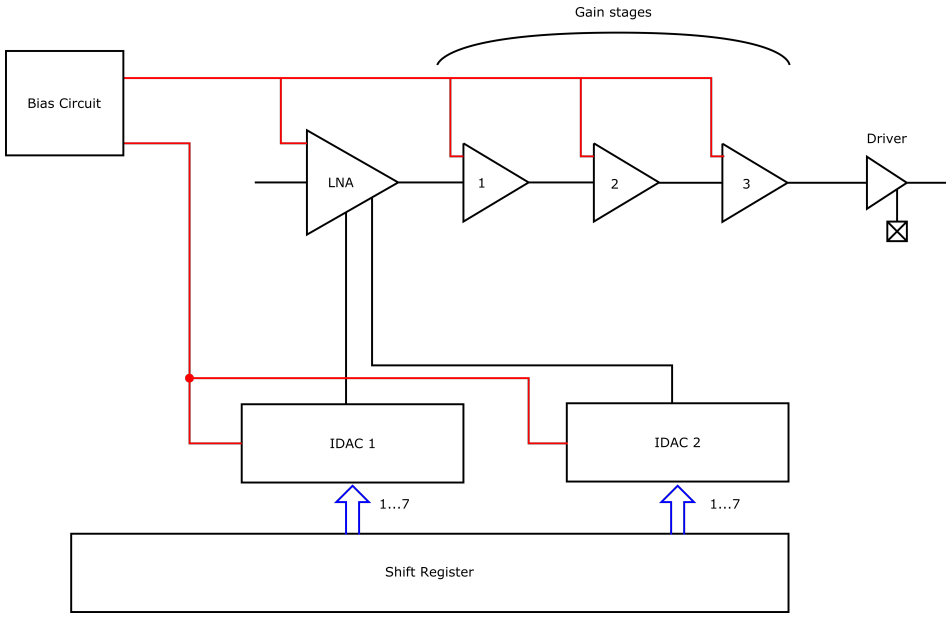

Figure 4.26: Schematic of full chip

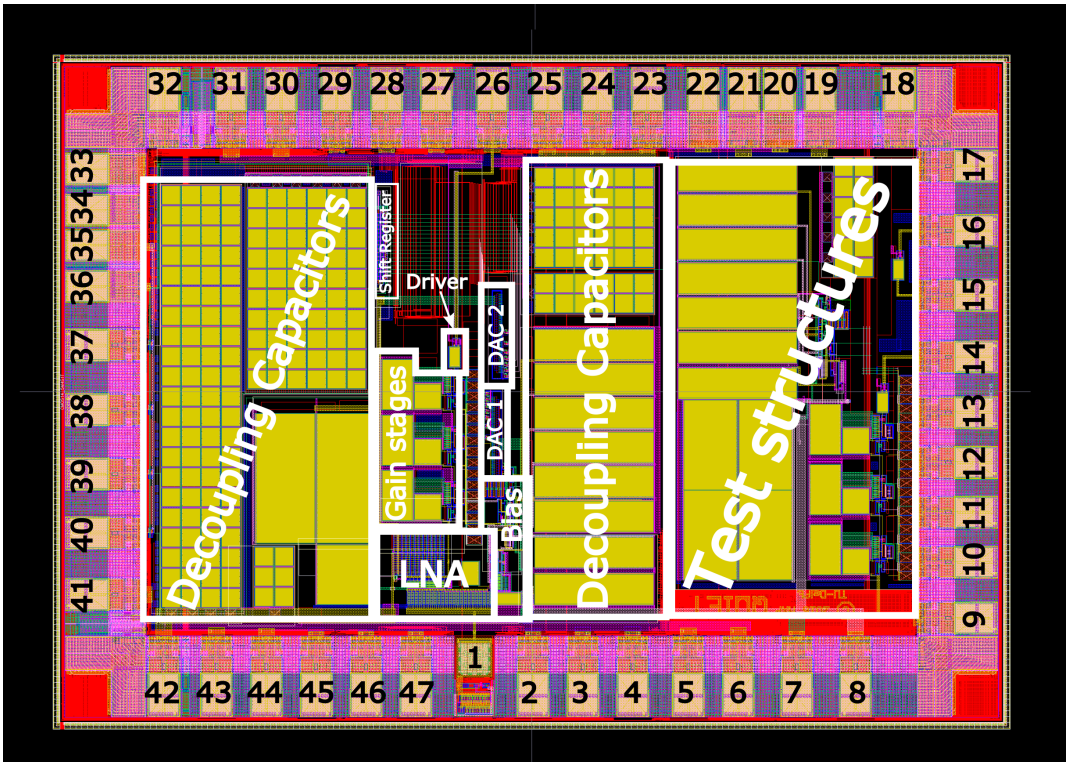

Figure 4.27: Layout of full chip

Many decoupling capacitors between supply lines and ground were added, to filter out noise and interference from the instruments, and to cancel the parasitic inductances in series with supply and ground at high frequencies. In particular, 358 pF for the supply of the gain stages, 537 pF for the supply of the LNA and 250 pF for the supply of the bias circuit. Further, 20 Ω resistor was added in series with these capacitors in order to decrease their quality factors and avoid ringing.

The pads are listed in tab 4.5 along with the corresponding net. Pad '1' was designed for low capacitance in order to minimize its effect on the bandwidth. Standard pads have a parasitic capacitance

Table 4.5: List of Pads

1	Input	**26**	Output
2	GND_LNA	**27**	GND_stages
3	GND_LNA	**28**	GND_stages
4	GND_LNA	**29**	GND_stages
5	Vdd_LNA	**30**	Vdd_stages
6	Vdd_LNA	**31**	Vdd_stages
7	I_{ext1}	**32**	Vdd_ ESD
8	Ext curr. source for bias	**33**	vddd
9	Input gain stages + driver test	**34**	gndd
10	GND_test	**35**	vddd
11	GND_test	**36**	gndd
12	GND_test	**37**	LOAD
13	GND_test	**38**	MOSI
14	Output gain stages + driver test f	**39**	CLK
15	Input driver test	**40**	MUX_sel
16	Output driver test	**41**	MISO
17	Vdd for test structures	**42**	I_{ext2}
18	Vdd for ESD	**43**	Vdd_LNA
19	GND_bias	**44**	Vdd_LNA
20	Vdd_bias	**45**	GND_LNA
21	GND_bias	**46**	GND_LNA
22	Vdd_bias	**47**	GND_LNA
23	GND_stages		
24	GND_stages		
25	GND_stages		

Table 4.6: Final performance of full chip

S21 [dB]	**Bandwidth (3-dB)**	**NF**	**Power**	**P/Qubit**
55 dB	740 MHz [a]	0.65 dB	75 mW	2.8 mW

[a] including bondwires' parasitics

to ground that amount to around 1 pF. This capacitance is due to the stack of metals (from Metal 1 to Metal 5) all connected in parallel. In the new pad, only Metal 5 and Metal 1 are connected in parallel. The other metals are left floating (not removed for mechanical stability). Furthermore, the metals above the ESD diodes were removed in order to further reduce the capacitance. In this way, the resulting capacitance amounted only to 300 fF, 30 % of the initial value. This value does not affect the bandwidth significantly, since the gate capacitance due to M_1 and M_{bottom} reaches 1.4 pF. The pads related to digital blocks, like 'DataIn', 'DataOut', 'LOAD', 'CLK', 'gndd' and 'vddd' were separated from the other pads by cutting the padring. Many ground pads were added, for the reason explained in the next section.

Test structures were also added, to double-check the correct functionality of each block. These structures include:

- Three gain stages plus driver
- Driver

Simulation S-parameters and noise figure after extraction of parasitics are shown in Fig. 4.28 and Fig. 4.29. The simulations include inductances from the bondwires which affect the overall performance. For more details, see section 4.4.

The overall performances are listed in table 4.6. The minimum achieved noise figure is 0.05 dB

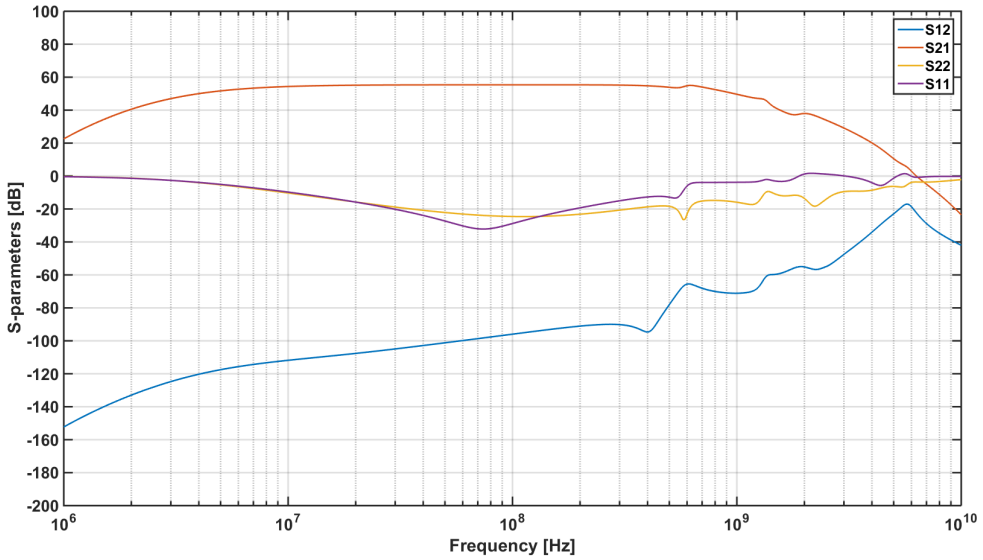

Figure 4.28: Post-Layout simulation of S-parameters, including bondwires' parasitics, explained in sec. 4.4

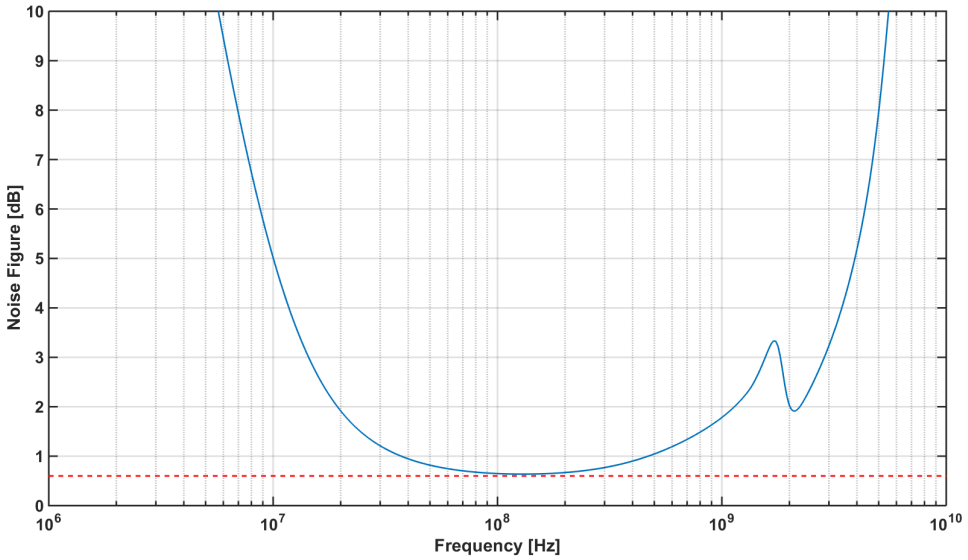

Figure 4.29: Post-Layout simulation of the noise figure, including bondwires' parasitics, explained in sec. 4.4

more than the target. This is due to the contribution of parasitic resistances and second stage. Nevertheless, the parasitic resistances are expected to reduce at cryogenic temperatures, due to increase in conductivity of the metals, and, therefore, the noise figure is expected to get closer to the target.

4.4. Testability issues

4.4.1. Grounding

Since the input signal is very weak and the gain is very large (55 dB), any coupling from the output back to the input may lead to performance degradation and even circuit instability. For instance, a little fluctuation on the output ground (due to large signals) can be comparable to the input signal. For this reason, the ground of the LNA was separated from the other grounds on chip. A schematic is shown in Fig. 4.30. Also the supplies have been separated, both for bondwires' issue and for power

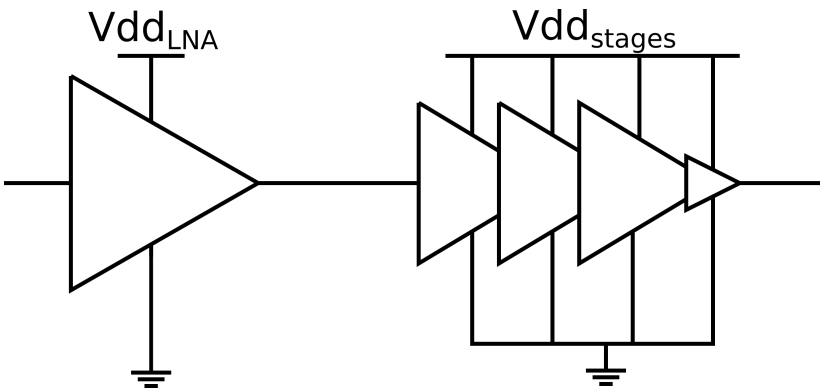

Figure 4.30: Grounding scheme

consumption measurements.

4.4.2. Bondwires

Two possibilities have been considered to interface the chip with the outside world: packaging or chip-on-board. Packaging was excluded because of the large parasitic capacitance of the pins. They can amount to more than 1 pF for a typical package like DIP-40 and they would jeopardize the input and output matching and eventually limit bandwidth. Chip-on-board, in which the bare die is bonded directly to the PCB was chosen. Nevertheless, even with chip-on-board, bondwires can not be avoided. These connections between silicon and PCB introduce large inductances (in the order of hundreds of pH) that can affect the behavior of the circuit. As a rule of thumb, since bondwire has an inductance of around 1 nH/mm, a typical 1-mm bondwire has an impedance of Z = j ωL = 6 Ωat 1 GHz. In Fig. 4.31 a model of the circuit including bondwires' parasitics is shown. Being the circuit single-ended, the

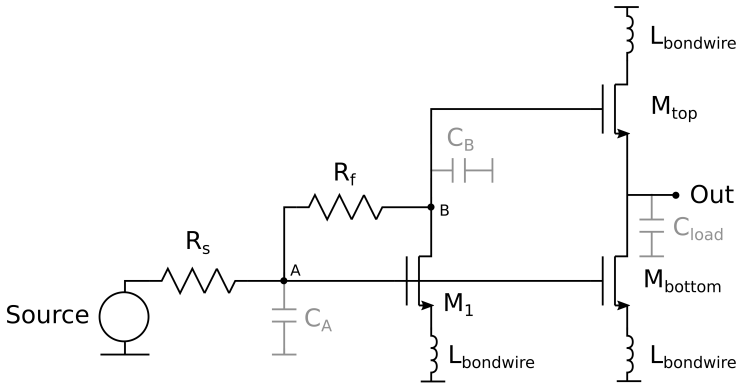

Figure 4.31: Small-signal circuit including model of bondwires

combination of a transistor with transconductance g_m and an inductor at its source is equivalent to a transistor with transconductance $\frac{g_m}{1+g_m Z}$. Since $g_{m\text{-bottom}}$ is around 350 mS the factor becomes ≈ 3 for an inductance of 1 nH at 1 GHz. This affects the bandwidth, the gain and the noise figure as shown in Fig. 4.32. In order to alleviate this issue, many grounds pads were added in the padring to allow multiple ground bondwires in parallel, thus decreasing the inductance value. In particular, 6 pads have been added for the ground of the LNA and 4 pads for the ground of the gain stages (for the pads see 4.5).

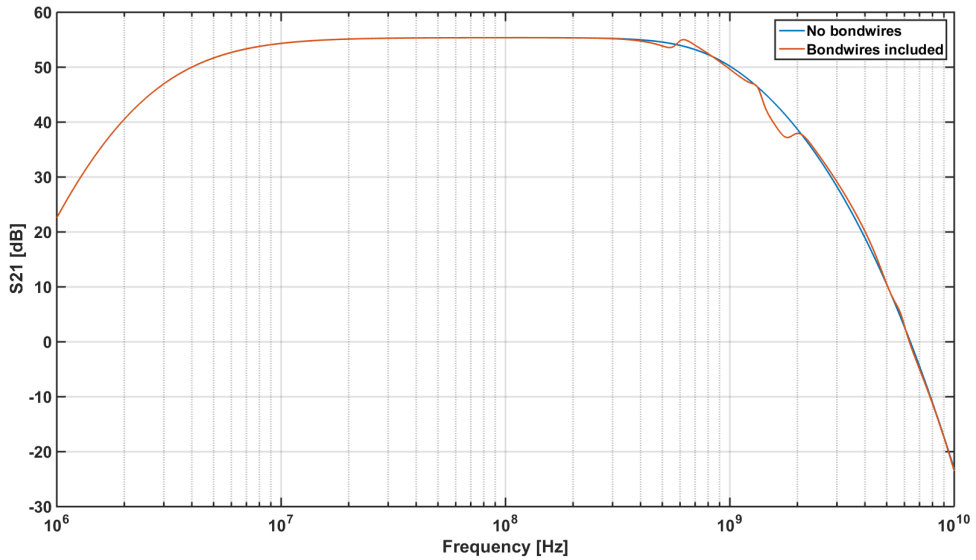

Figure 4.32: Comparison of extracted S21 including bondwires' inductances with S21 without bondwires. The gain looses is flatness at high frequencies and the bandwidth drops by almost 50 MHz

5

Measurements

In this chapter, the characterization of the chip will be presented. First, the setup will be described and secondly the results shown. Then, the results of both room temperature (RT) and liquid-helium temperature (LHT) measurements will be presented and discussed.

The micrograph of the fabricated chip (technology SSMC 0.16 µm), is shown in Fig. 5.1 where the core circuit is highlighted and overall dimensions shown.

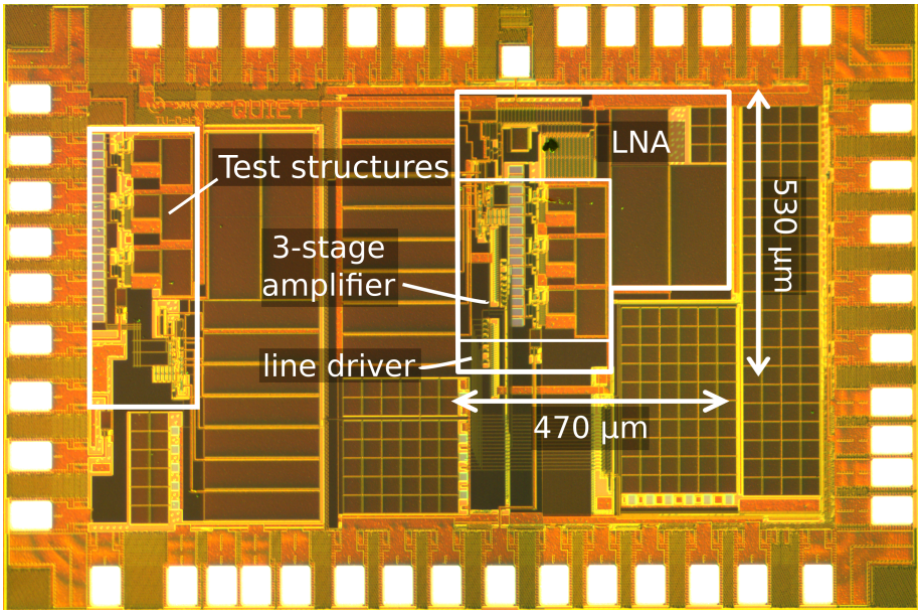

Figure 5.1: Micrograph of the chip. Dimensions of the core circuit are shown.

5.1. Printed Circuit Board

As described in chapter 4, the chip was not packaged but bonded directly on a printed circuit board (PCB). The design of the PCB, shown in Fig. 5.2, had to comply with the chip specifications about bondwires and grounding, as discussed in chapter 4. To reduce the effect of the bondwires, the chip was placed on a grounded metal plate, whose dimensions exceed the dimensions of the chip by 400 µm on each side. This enabled the use of very short bondwires from the GND pads to this metal plate, thereby reducing the equivalent parasitic inductance of the connections. Moreover, thick bondwires were used (diameter = 25 µm), to further alleviate this issue. To address the grounding, the bottom layer of the PCB was split in two main ground planes (orange and blue squares in Fig. 5.2). The analog ground of the LNA was connected on one side (orange), while the ground of all the other blocks (bias circuit, gain stages and driver, digital ground) was routed from the chip on the other plane

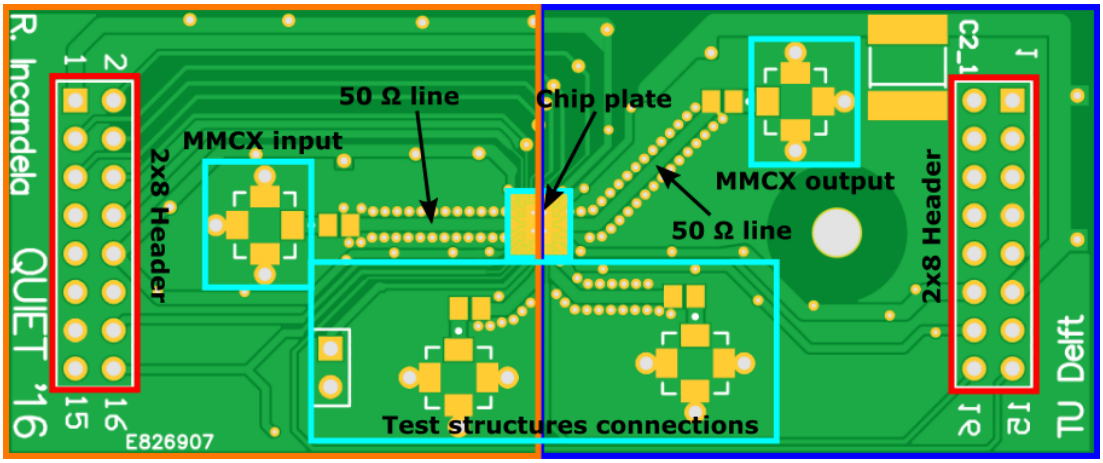

Figure 5.2: PCB layout

(blue). This prevents any large signal to affect the input node. The two ground planes were then wired and connected together to the main ground on one of the bench-top supply, in order to equalize the reference node.

The input and output signals were routed onto a 50-Ω coplanar waveguide (CPW), for input and output matching The waveguide is connected to MMCX connectors through AC-coupling capacitors (470 pF). All the other signals were routed from the chip to two 8x2 headers, used to connect all cables.

Another important feature of this PCB is the components choice. The most critical components were the decoupling capacitors: the purpose of these devices is to act as local energy storage for high-frequency current demands and as shunting path for noise superimposed to the supply. According to literature, [41], the only dielectric robust to such wide range of temperatures (4 K - 300 K) is of the type C0G or NP0. All other types of capacitors reduce their nominal capacitance value by more than 70 %. Most of the capacitors were placed on the bottom layer of the board and connected through vias to their respective traces. As a rule of thumb, decoupling capacitors must be placed such that the smallest value is close to the chip while capacitors with larger values are placed in parallel and farther, in order to optimally exploit the self-resonance of the capacitors. The smallest capacitor used for all supplies is 470 pF, in parallel with capacitors of 4.7 nF, 47 nF and 470 nF. Another capacitor of 4.7 µF and type X7R was added in parallel even if it is supposed to fail at LHT. This could at least help at room temperature to filter out low-frequency interferences from the supplies. Finally, the sizes of the PCB are 30mm x 70 mm due to space constraints as explained in Sec. 5.2.2.

5.2. Setups

5.2.1. Room-temperature setup

Even though the circuit was meant to work at cryogenic temperatures, a first test at room temperature was performed to check the functionality and compare the performance to simulation, where the models are certified. In this section, the setup is presented and the performed measurements described.

As mentioned in Sec. 5.1, all the DC signals were wired to the two 2x8 headers. In Fig. 5.3 the top-level schematic is shown. In total, four 1.8 V power supplies were needed (Vdd_{LNA}, Vdd_{stages}, Vdd_{bias} and $V_{digital}$) and four external current sources (I_{ext1}, I_{ext2}, I_{bias} and I_{driver}). The supplies were provided by a supply board which consists of 10 LDO regulators. The supply board was then connected to a RIGOL DP832A analog supply. The bias currents were provided by Keithley 2636B SMUs.

The digital signals, $Data_{in}$, $Data_{out}$, CLK and LOAD were provided by an FPGA (Artix 7) board (Nexys 4) connected to the left header through a voltage divider, to adapt the 3.3 V levels to 1.8 V. The clock frequency was set to 450 kHz. MUX_{SEL}, which decides if the shift register is in default state or not, was connected to fixed voltage supply. The bit configuration was sent to the FPGA through Matlab from a common laptop.

For S-parameters measurement, the input and output were connected to a vector network analyzer (Hewlett Packard 85047A). The VNA was calibrated with the four standards: *open*, *short*, *load* and *through* before performing RT measurements. The board is not included in the calibration and, there-

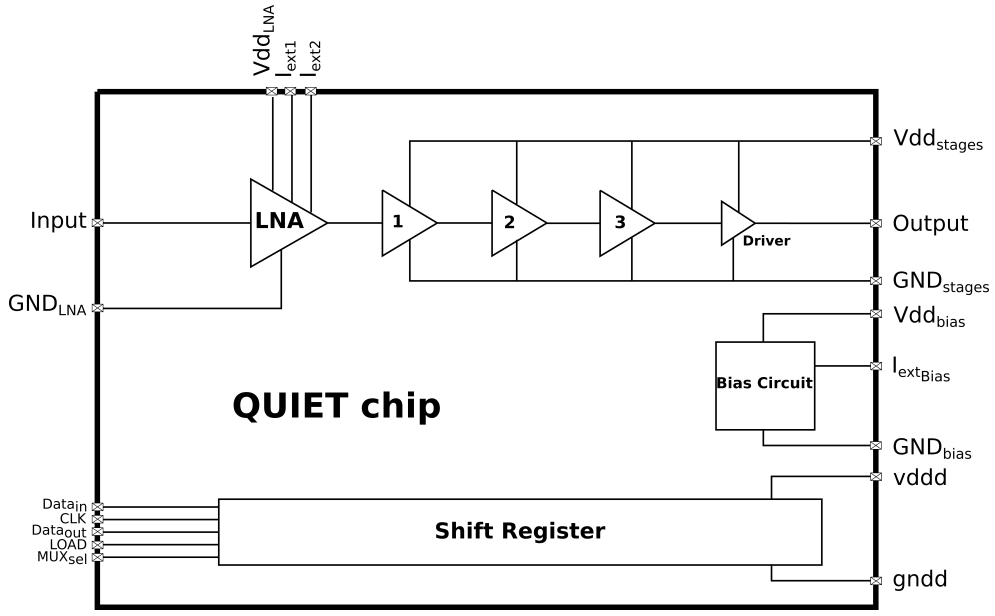

Figure 5.3: Top-level schematic of chip

fore, any error in the designed CPW can affect the performance of the entire chip. The calibration was not repeated at 4 K because of the unavailability of standards (50 Ω, open, short, through) that can operate at 4 K.

Furthermore, the same measurements were performed two times, first with short cables (around 30 cm) from instruments to PCB and then with long cables (around 1.5 m, same length as the pipe described in section 5.2.2). This was to analyze the effect of the setup on the performance, since for the cryogenic setup, long cables are needed to connect the PCB dipped in liquid Helium to room temperature instruments, as described in next section 5.2.2. All the cables are of type coaxial with 50-Ω characteristic impedance.

The final setup is shown in Fig. 5.4. The blue lines represent the ground connections, while the red lines represent the supply and bias current connections. All the grounds of the Keithley, FPGA and supply board are star-connected to the RIGOL supply. This connection is not shown to ease the readability of the figure.

5.2.2. Cryogenic setup

In order to characterize the chip at cryogenic temperature, the setup in Fig. 5.5 was used. It consists of a liquid Helium dewar and a long metal pipe (1.5 m) inserted into the vessel. The PCB is placed at the bottom of the pipe, screwed to the metal for mechanical stability. All the cables are placed in the pipe, in order to connect the printed circuit board (PCB), on one side, to a metal box on the other side. In the metal box at the top, BNC connectors link inner cables to outer cables used to connect to bench-top instruments. The same instruments as in the RT measurements were used here.

5.3. Results

5.3.1. Measurement issues

The first test was performed at RT and with the chip in default state, which means that the shift-register was disabled by the 'default' signal and all the bias currents were supplied externally. When connected to the oscilloscope, it was noticed that all the nodes were oscillating. The oscillating frequency was relatively low (approximately 10 MHz), thus leading to believe that the oscillation did not originate from the chip but more likely from an external loop in the measurement setup. In particular, if a discrete capacitor was connected from pin I_{ext1} to ground, the oscillation frequency almost halved and the amplitude of oscillation decreased. For this reason, the external current sources were discarded for all the next measurements and all the bias currents were generated on chip and programmed by the FPGA, except for the bias current of the driver that cannot be generated internally.

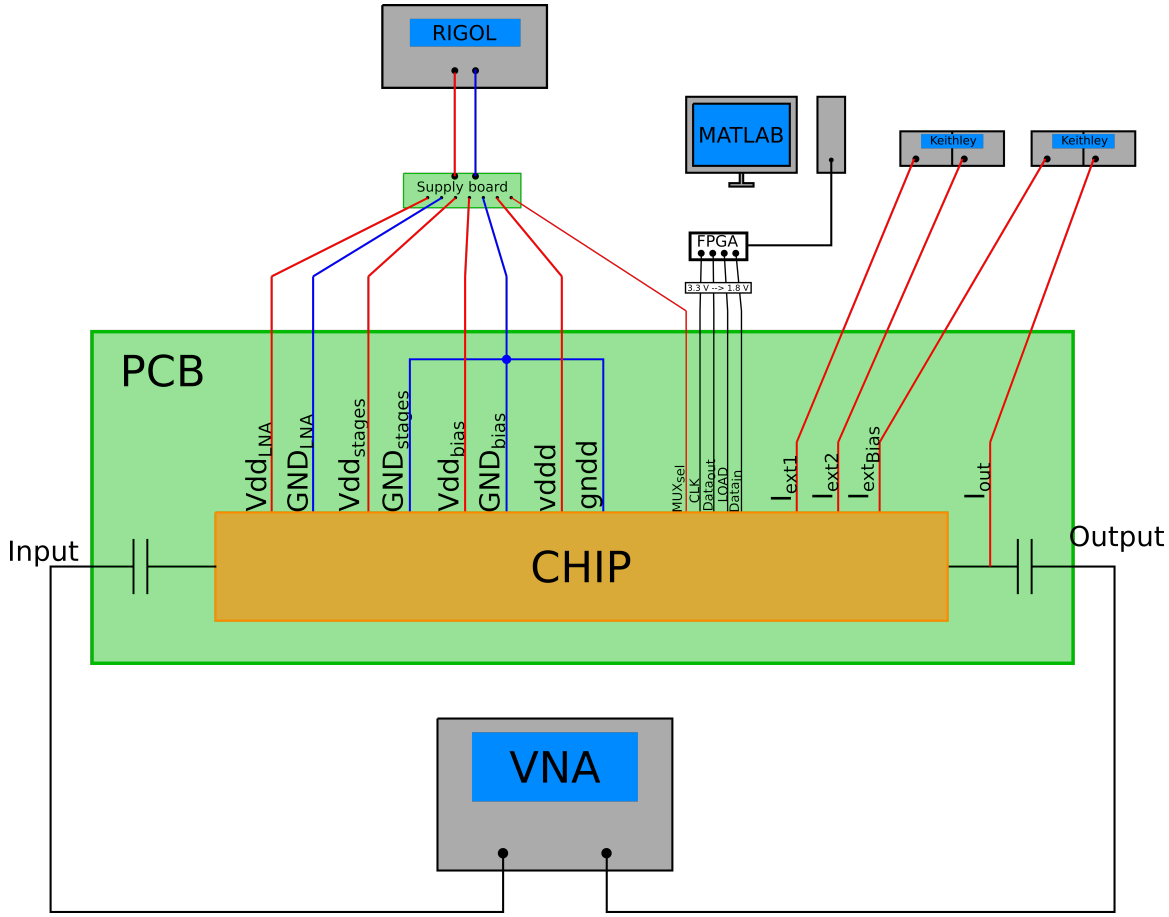

Figure 5.4: Measurement setup.

5.3.2. Results

In order to find the optimal operation point, manual tuning was performed for the current values to find the point where S11 was minimum and gain maximum. Finally, the shift-register was configured for internal bias currents and the decimal value sent to I_{DAC1} equal to 12 and equal to 90 for I_{DAC2}. Also the bias current of the driver was manually tuned to find the best spot for S22 and finally set to 4 mA. For cryogenic measurements the current settings were the same as at RT except for I_{DAC1} whose decimal value in the shift-register was set to 6. The amplifier draws 32 mA from Vdd_{LNA}, 20 mA from Vdd_{stages} and 2 mA from Vdd_{bias}, for an overall power consumption of 97 mW at LHT. At RT the bias circuit draws 1 mA and the LNA 28 mA. The stages consumes almost the same power and the overall power consumption is 84.6 mW. From RT simulation a power of 73.8 mW was expected. This variation of 15 % can be due to process spread. From LHT, on the other hand, a power consumption of 75 mW was expected, but in this case, the mismatch is more likely due to the non-exact model used for LHT simulations.

The measured and simulated S-parameters are shown in Fig. 5.6,5.7,5.8, 4.25.

5.3.3. Observations and discussion

From the measurement results, the following observations can be made:

S11 From Fig. 5.6 it can be seen that S11 is close to -5 dB at RT, while it shows large peaks at LHT. The average value of S11 at LHT is, anyway, lower than at RT, showing an improvement of more than 2 dB. The first important observation is about the difference results when the amplifier is measured with short-cables setup (red dots in Fig. 5.6) and long cables (blue dots in Fig. 5.6). While with short cables S11 is almost flat over the whole bandwidth (20 MHz - 650 MHz), with long cables it shows notches and peaks of 5 dB. These peaks become even worse when the amplifier is cooled down. This ripple can

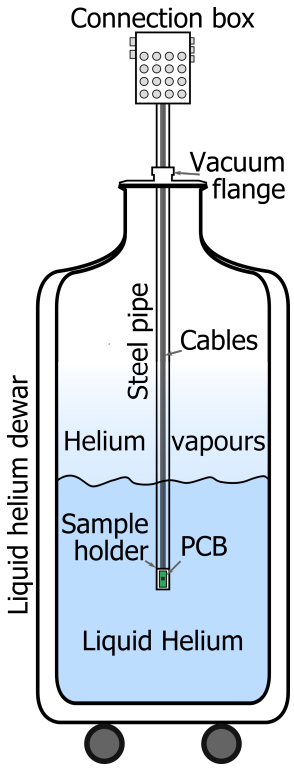

Figure 5.5: Cryogenic setup.

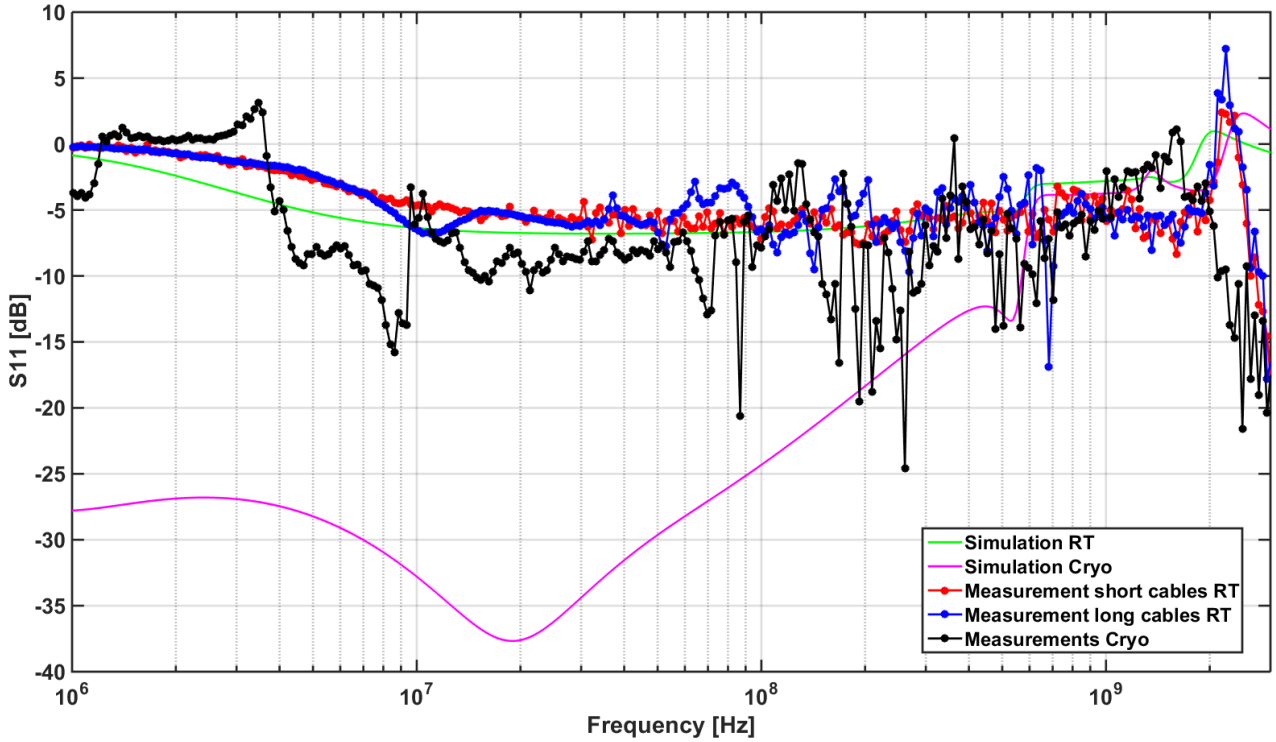

Figure 5.6: S11. The figure compares RT simulation, LHT simulation, short and long cables measurements and LHT measurements

be due to multiple reflections in the input coaxial cable caused by mismatch between the line and the waveguides on the board, that were not considered in the calibration. A hypothesis on the deterioration

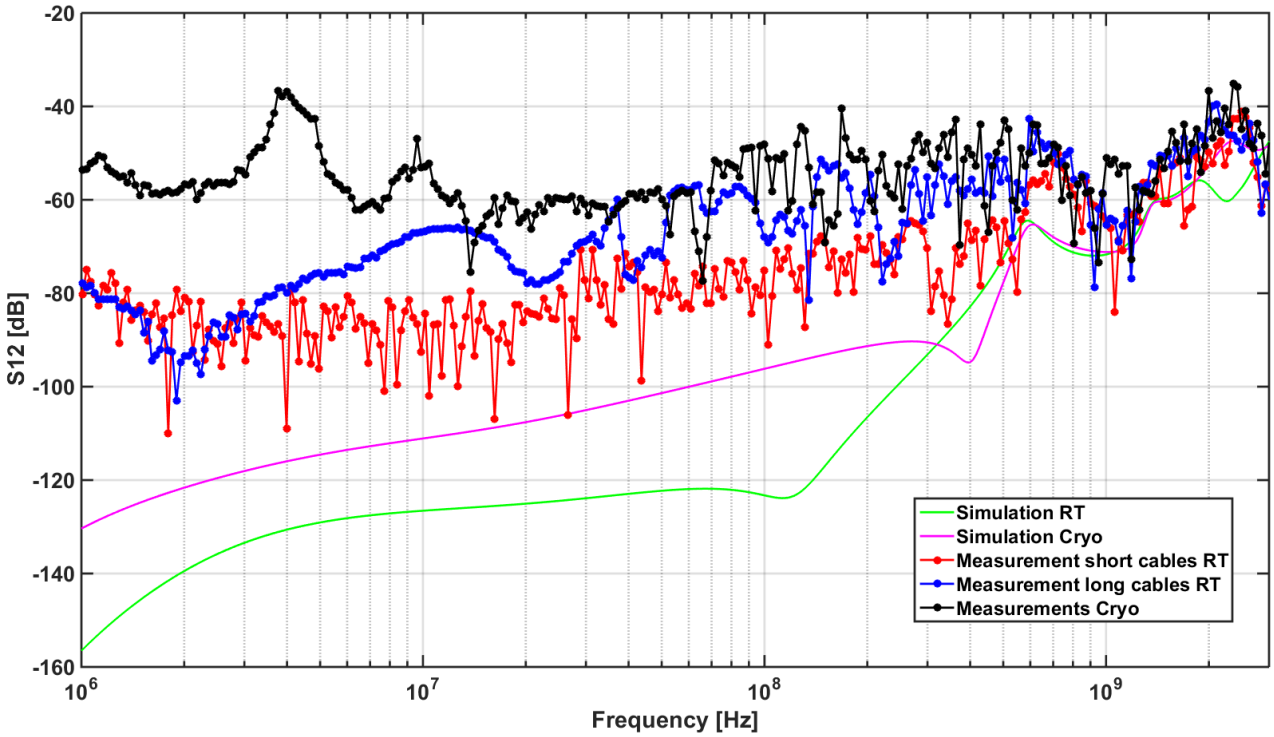

Figure 5.7: S12. The figure compares RT simulation, LHT simulation, short and long cables measurements and LHT measurements

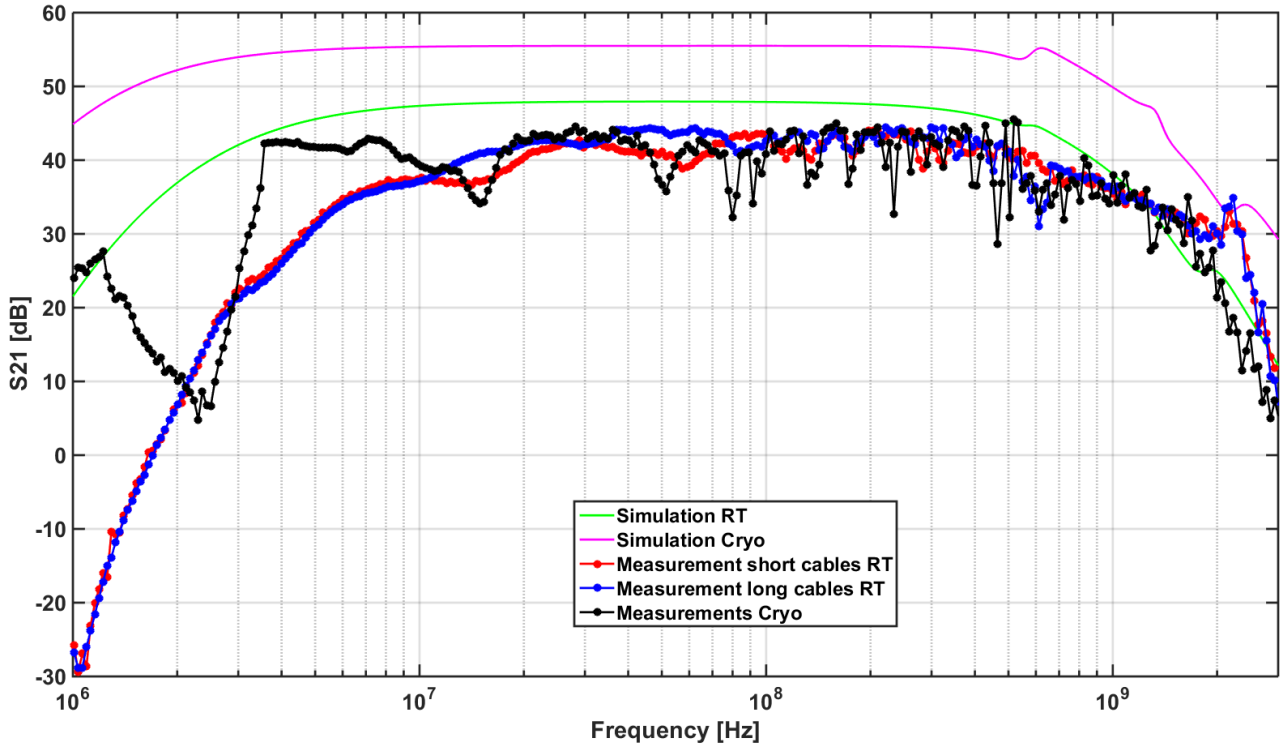

Figure 5.8: S21. The figure compares RT simulation, LHT simulation, short and long cables measurements and LHT measurements

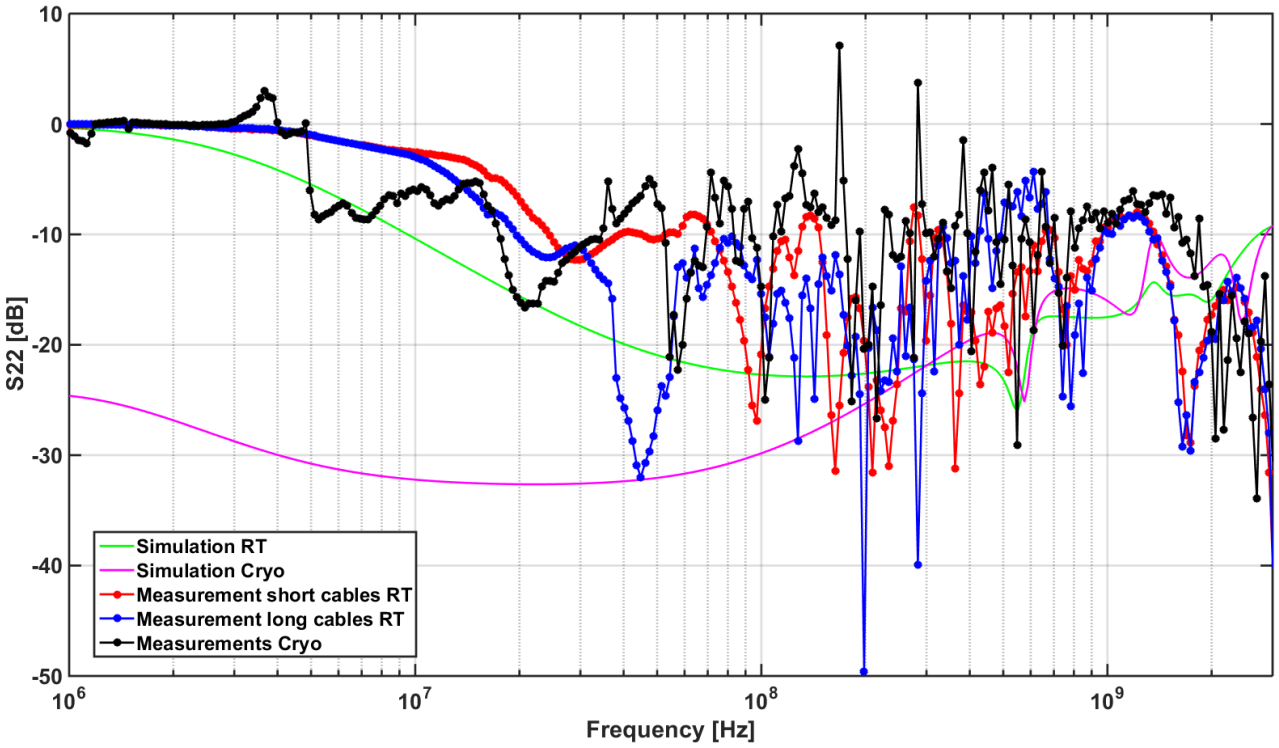

Figure 5.9: S22. The figure compares RT simulation, LHT simulation, short and long cables measurements and LHT measurements

at LHT could be the temperature dependence of the losses of the coaxial cables and the degradation of the discrete capacitors on the PCB for decoupling of the supplies. Moreover, it has to be recalled that no calibration was performed at 4 K but the RT calibration data were used to measure the S-parameters at LHT. This can add large errors on top of the measurements. The RT measurements show good agreement with RT simulations performed with the standard libraries provided by the foundry. On the other hand, the cryogenic model shows large mismatch with the cryogenic measurement. As it was mentioned in chapter 4, a coarser model was adopted for the design, different from the one described in chapter 2. This could be a reason for this large mismatch. At around 2.2 GHz, S11 becomes larger than 0 dB. This is not possible, unless other sources of power are coupled to the input and read by the VNA as reflected power. This peak is observed also in simulation and disappears if the inductance of the bondwires is removed. This clearly shows the limitations of the bondwires, explained in chapter 4. At LHT this peak moves to slightly lower frequency: this could be due to a small change in the inductance when the temperature is lowered. Anyway it is out of the band of interest, therefore it will not cause any significant problem.

S12 S12, which represents the input response for a signal applied at the output of the amplifier and it is shown in Fig. 5.7, results to be less than -60 dB in the bandwidth 20 MHz - 650 MHz for the short-cables setup, less than -55 dB with long cables and below -40 dB for the cryogenic measurement. The large mismatch with the simulations is mainly due to the fact that the simulator takes in account only possible capacitive feedback between input and output while in reality also coupling through the substrate plays an important role. The substrate coupling, on the other hand, cannot be easily simulated with a simple 'sp' simulation. An observation must be made here: at LHT the substrate is expected to be frozen and lowly conductive. An improvement in S12 would be expected if the main coupling came from there. Nevertheless a degradation of S12 is observed from 1 MHz to 30 MHz, although it is out of the band of interest. Also for S12 peaking is observed and the same hypothesis as for S11 are valid in this case.

S21 S21, that represents the gain of the amplifier, shows an in-band gain of around 42 dB (Fig. 5.8). The bandwidth of information is from 20 MHz to 650 MHz defined as maximum gain minus 3-dB on each side. At LHT, a very sharp step is observed at 3.5 MHz. The reason of this cannot only be due to the high-pass filters between the stages and it is not clear yet. A slower response should have been expected as for the RT measurements. If compared to simulations, the three measurements agree with the RT model while the cryogenic model shows larger gain. The in-band gain decreased 4 dB after fabrication. This could be explained by recalling that the amplifier consists of many stages and, therefore, a little mismatch in gain of each of them produces a large overall difference (since the error is multiplied). Nevertheless, 4 dB is an acceptable error, since some margin was considered while designing. The high-pass filters made by the AC-coupling capacitors between each stage moved to slightly higher frequency, from 3 MHz in simulation to around 20 MHz in measurement. This can easily be due to process spread. High peaks and valleys are observed in the cryogenic measurement as in the other S-parameters.

S22 S22 quantifies the quality of the output matching. In Fig. 5.9 measurements and simulations are compared. S22 can be tuned with the bias current (I_{out}) of the driver described in chapter 4. An optimum current of 4 mA was found at RT. At LHT the output current was kept the same because no large improvements was found with changing it. This was probably due to the large peaks that hid the real improvements. Large peaking is observed for all the measured curves, at RT and at LHT. The cause of this large variations, even at RT is not yet clear. The SMU that generates the current is a possible factor of degradation. In fact, it was found that the instrument injects power directly to the output of the amplifier even if the current is switched off. By measuring the output spectrum of the amplifier without any input, large peaks up to -55 dB were observed, even if the current from the SMU was switched off. Only after switching off the SMU the output spectrum (with no signal at the input) returned flat. This could explain some of the wrong behavior of S22 but no exact evidence was found.

5.4. Noise measurement

Because of time constraints, the noise performance of the amplifier has not been evaluated. Nevertheless, in this section a short description of how the noise could be evaluated will be given. As mentioned in 4.3.5, the noise floor at the input of the LNA is estimated to be -186 dBm/Hz. Considering 42 dB of gain (S21 at LHT), at the output of the amplifier -144 dBm/Hz should be observed. In order to emulate the noise of the 50-Ω source resistance, a 50-Ω resistor can be inserted at the input of the amplifier. The idea is to measure the noise floor at the output of the amplifier with a spectrum analyzer and manually retrieve the noise figure. The main limitation of this measurement, also called Gain-method, is the temperature spread of the 50-Ω resistor. The assumption that it holds the same value and that its noise scales linearly with temperature must be verified before performing the measurement. An alternative method is based on the so-called Y-factor. It consists of injecting two different levels of noise at the input, with two different noise sources, kT_1B and kT_2B where T is the noise temperature of the source. The output power of the amplifier is then computed for these two input noise levels:

$$\begin{cases} P_1 = Gk(T_1 + T_{amp})B \\ P_2 = Gk(T_2 + T_{amp})B \end{cases} \tag{5.1}$$

where G is the gain, B is the bandwidth and T_{amp} is the noise temperature of the amplifier. The output power is then converted in noise temperature T_{out}:

$$\begin{cases} T_{out1} = G(T_1 + T_{amp}) \\ T_{out2} = G(T_2 + T_{amp}) \end{cases} \tag{5.2}$$

If plotted as a function of the noise temperature T, the two values of T_{out} lay on a line whose slope is G and whose x-intercept is $-T_{amp}$, as shown in Fig. 5.10. This approach suffers from the uncertainty on the noise sources and on the input matching. In fact, these noise sources must provide 50 Ω input matching and this is known to be an important issue [42].

In general, both approaches must be investigated and the noise figure obtained with reasonable accuracy. Being the noise figure extremely small (0.009 dB), every minimum source of error will not be negligible.

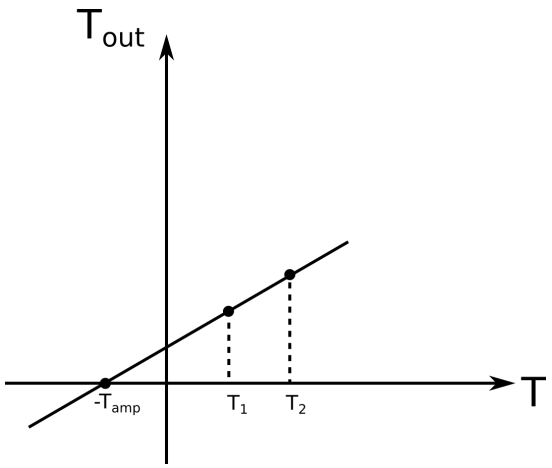

Figure 5.10: Y-factor method

5.5. Conclusions

In conclusion, the chip proved to be functional both at RT and LHT. The performance, on the other hand, must be improved and in particular the peaks must be better understood and, if possible, removed. This is important in order to discriminate the sources of error from the setup or from the chip and to obtain new insights on both. Noise performance was not evaluated because of time contraints. The final performance is listed in table 5.1.

Table 5.1: Final comparison between post-layout simulations with adapted model to LHT and performance after testing

	S21 [dB]	**Bandwidth**	**NF**	**Power**	**P/Qubit**
Post-layout	55	740 MHz	0.65 dB	75 mW	2.8 mW
Measurement	42	650 MHz	Not measured	97 mW	4.3 mW

6

Conclusions and Future works

6.1. Conclusions

In this work, the design of the first cryoCMOS LNA to be interfaced with spin qubits was presented. To design the amplifier and ensure good level of complexity, MOS11 model was adapted to cryogenic temperature by modifying certain crucial parameters. To do so, a thorough study of cryogenic behavior of CMOS devices at 4 K was performed by comparing prior literature with characterization performed in our group, as described in Chapter 3.

From this work, the following observations could be made:

- Kink effect, known as one of the most hampering anomaly for cryoCMOS, is not present in technology nodes with feature size smaller or equal to 0.16 µm.
- Subthreshold slope and mobility are strongly enhanced, bringing advantages and possibilities for cryoCMOS not yet explored.
- Compact modeling is possible with existing models that can cover a very large range of temperatures, modifying certain parameters and/or augmenting the device with a simple resistor to model the increase in substrate resistance due to freeze out (for thick-oxide transistors).
- The RF-reflectometry setup was analyzed from the electrical point of view, enabling the derivation of certain specifications required for the design of the amplifier.
- CMOS functionality has been further confirmed at cryogenic temperatures with the design of an LNA and the circuitry around it (shift register, DAC, self-biased current generator)
- The amplifier showed a wide bandwidth and large gain at 4 K
- The experimental setup is still a strong limitation for performing cryogenic measurements

State-of-the-art readout of spin qubits is performed with bulky discrete components. With this work, we laid the basis for the integration of these complex systems, a step needed towards the actualization of quantum computers.

6.2. Future work

Short term A complete characterization of the chip has not been performed yet. Therefore, many activities are to be completed:

- First of all, the notches in the transfer function need to be removed, or at least their cause must be understood. So far, a hypothesis on the setup was made but no experimental confirmation were reported.
- Secondly, the noise performance must be evaluated.

- The test structures need also to be tested to discriminate the sources of errors: do they come from the LNA or from the additional gain stages or the driver?
- Finally, robustness can be tested by measuring how the amplifier respond to various stress: changes in supply voltage, changes in bias current, sweep versus temperatures etc.
- When the amplifier is fully characterized, testing it along with a real QPC must be done. This will enable to confirm the hypothesis and assumptions made in this thesis, as well as finding new insights to improve the electrical model of the QPC. This is a key point for properly engineering the spin-qubit readout.

For what the modelling and characterization of cryoCMOS is concerned, a lot of work needs still to be done:

- First of all, dynamic characterization and modelling need to be done. This will enable reliable transient simulations.
- Secondly, mismatch need to be evaluated at such low temperatures. Literature on this topic, [43], reports a drastic degradation in matching of transistors at low temperatures but no exhaustive explanation was provided.
- Flicker noise needs to be modelled. This will enable to take in consideration also low-frequency noise and design of accurate low-frequency circuits.

Long term In the future, along with an improved quality of qubits and their charge sensors, a full optimized readout of spin qubit should be designed. The fundamental limitations of cryoCMOS should also be investigated because, until now, the boundary between advantages and disadvantages is not yet clear: could we really defeat thermal noise by going to lower temperatures? Up to which temperature cryoCMOS devices are still functional and why do they fail if we decrease the temperature further? Up to which temperature cryoCMOS brings advantages and at which temperature disadvantages overcome the benefits? All these question must be answered if we want to make cryoCMOS the technology quantum computers will rely on.

Bibliography

[1] R. P. Feynman, *Simulating physics with computers,* Int'l J. Theoretical Physics **21**, 467 (1982).

[2] K. M. Svore and M. Troyer, *The quantum future of computation,* Computer **49**, 21 (2016).

[3] P. J. J. O'Malley, R. Babbush, I. D. Kivlichan, J. Romero, J. R. McClean, R. Barends, J. Kelly, P. Roushan, A. Tranter, N. Ding, B. Campbell, Y. Chen, Z. Chen, B. Chiaro, A. Dunsworth, A. G. Fowler, E. Jeffrey, E. Lucero, A. Megrant, J. Y. Mutus, M. Neeley, C. Neill, C. Quintana, D. Sank, A. Vainsencher, J. Wenner, T. C. White, P. V. Coveney, P. J. Love, H. Neven, A. Aspuru-Guzik, and J. M. Martinis, *Scalable quantum simulation of molecular energies,* Phys. Rev. X **6**, 031007 (2016).

[4] D. P. DiVincenzo, *The physical implementation of quantum computation,* Fortschritte der Physik **48**, 771 (2000).

[5] R. V. Meter and S. J. Devitt, *The path to scalable distributed quantum computing,* Computer **49**, 31 (2016).

[6] R. Van Meter and S. J. Devitt, *Local and Distributed Quantum Computation,* ArXiv e-prints (2016), arXiv:1605.06951 [quant-ph] .

[7] R. Hanson, L. P. Kouwenhoven, J. R. Petta, S. Tarucha, and L. M. K. Vandersypen, *Spins in few-electron quantum dots,* Rev. Mod. Phys. **79**, 1217 (2007).

[8] R. Hanson, *Electron spins in semiconductor quantum dots*, Ph.D. thesis, TU Delft (2005).

[9] H. van Houten and C. Beenakker, *Quantum point contacts,* Physics Today **49** (1996), 10.1063/1.881503.

[10] Wikipedia, *Electron paramagnetic resonance,* .

[11] J. M. Elzerman, R. Hanson, L. H. Willems van Beveren, B. Witkamp, L. M. K. Vandersypen, and L. P. Kouwenhoven, *Single-shot read-out of an individual electron spin in a quantum dot,* Nature **430**, 431 (2004).

[12] B. Brun, F. Martins, S. Faniel, B. Hackens, A. Cavanna, C. Ulysse, A. Ouerghi, U. Gennser, D. Mailly, P. Simon, S. Huant, V. Bayot, M. Sanquer, and H. Sellier, *Electron phase shift at the zero-bias anomaly of quantum point contacts,* Phys. Rev. Lett. **116**, 136801 (2016).

[13] *Cryogenic SiGe Low Noise Amplifier*, California Institute of Technology (2013).

[14] L. M. K. Vandersypen, J. M. Elzerman, R. N. Schouten, L. H. Willems van Beveren, R. Hanson, and L. P. Kouwenhoven, *Real-time detection of single-electron tunneling using a quantum point contact,* Applied Physics Letters **85**, 4394 (2004).

[15] J. van Oven, *Radio-frequency-reflectometry measurements on quantum dots using nbtin spiral inductors,* (2013).

[16] T. Müller, T. Choi, S. Hellmüller, K. Ensslin, T. Ihn, and S. Schön, *A circuit analysis of an in situ tunable radio-frequency quantum point contact,* Review of Scientific Instruments **84**, 083902 (2013), http://dx.doi.org/10.1063/1.4817306.

[17] A. G. Fowler, M. Mariantoni, J. M. Martinis, and A. N. Cleland, *Surface codes: Towards practical large-scale quantum computation,* Phys. Rev. A **86**, 032324 (2012).

[18] J. M. Hornibrook, J. I. Colless, A. C. Mahoney, X. G. Croot, S. Blanvillain, H. Lu, A. C. Gossard, and D. J. Reilly, *Frequency multiplexing for readout of spin qubits,* Applied Physics Letters **104**, 103108 (2014), http://dx.doi.org/10.1063/1.4868107.

[19] D. A. Neamen, *Semiconductor Physics and Devices*, edited by McGraw-Hill.

[20] E. A. Gutierrez-D, M. J. Deen, and C. L. Claeys, *Low Temperature Electronics: Physics, Devices, Circuits, and Applications*, edited by A. Press.

[21] S. Wang, *Fundamentals of semiconductor theory and device physiscs*, edited by P. H. C. Div.

[22] Y. Feng, P. Zhou, H. Liu, J. Sun, and T. Jiang, *Characterization and modelling of mosfet operating at cryogenic temperature for hybrid superconductor-cmos circuits,* Semiconductor Science and Technology **19**, 1381 (2004).

[23] T. C. Carusone, D. A. Johns, and K. W. Martin, *Analog Integrated Circuit Design*, edited by J. Wiley and Sons.

[24] F. Balestra, L. Audaire, and C. Lucas, *Influence of substrate freeze-out on the characteristics of mos transistors at very low temperatures,* Solid-State Electronics **30**, 321 (1987).

[25] B. Dierickx, L. Warmerdam, E. R. Simoen, J. Vermeiren, and C. Claeys, *Model for hysteresis and kink behavior of mos transistors operating at 4.2 k,* IEEE Transactions on Electron Devices **35**, 1120 (1988).

[26] Y. Creten, P. Merken, W. Sansen, R. P. Mertens, and C. V. Hoof, *An 8-bit flash analog-to-digital converter in standard cmos technology functional from 4.2 k to 300 k,* IEEE Journal of Solid-State Circuits **44**, 2019 (2009).

[27] E. Simoen, B. Dierickx, L. Deferm, C. Claeys, and G. Declerck, *The charge transport in a silicon resistor at liquid□helium temperatures,* Journal of Applied Physics **68** (1990).

[28] L. Deferm, E. Simoen, and C. Claeys, *The importance of the internal bulk-source potential on the low temperature kink in nmosts,* IEEE Transactions on Electron Devices **38**, 1459 (1991).

[29] S. H. Hong, G. B. Choi, R. H. Baek, H. S. Kang, S. W. Jung, and Y. H. Jeong, *Low-temperature performance of nanoscale mosfet for deep-space rf applications,* IEEE Electron Device Letters **29**, 775 (2008).

[30] A. Siligaris, G. Pailloncy, S. Delcourt, R. Valentin, S. Lepilliet, F. Danneville, D. Gloria, and G. Dambrine, *High-frequency and noise performances of 65-nm mosfet at liquid nitrogen temperature,* IEEE Transactions on Electron Devices **53**, 1902 (2006).

[31] A. H. Coskun and J. C. Bardin, *Cryogenic small-signal and noise performance of 32nm soi cmos,* in *2014 IEEE MTT-S International Microwave Symposium (IMS2014)* (2014) pp. 1–4.

[32] G. D. Geronimo, A. D'Andragora, S. Li, N. Nambiar, S. Rescia, E. Vernon, H. Chen, F. Lanni, D. Makowiecki, V. Radeka, C. Thorn, and B. Yu, *Front-end asic for a liquid argon tpc,* IEEE Transactions on Nuclear Science **58**, 1376 (2011).

[33] J. Chang, A. A. Abidi, and C. R. Viswanathan, *Flicker noise in cmos transistors from subthreshold to strong inversion at various temperatures,* IEEE Transactions on Electron Devices **41**, 1965 (1994).

[34] Y. P. Tsividis, *Operation and Modeling of the MOS Transistor*, edited by McGraw-Hill.

[35] X. Li, W. Wu, G. Gildenblat, G. D. J. Smit, A. J. Scholten, D. B. M. Klaassen, and R. Langevelde, *PSP 102.3*.

[36] R. Langevelde, A. J. Scholten, and D. B. M. Klaassen, *Physical background of MOS Model 11*.

[37] L. Song, *Cryogenic characterization and modeling of nanometer cmos for quantum computing applications,* (2016).

[38] B. Razavi, *RF Microelectronics*, edited by P. Hall.

[39] F. Bruccoleri, E. A. M. Klumperink, and B. Nauta, *Wide-band cmos low-noise amplifier exploiting thermal noise canceling,* IEEE Journal of Solid-State Circuits **39**, 275 (2004).

[40] F. Bruccoleri, E. Klumperink, and B. Nauta, *Wideband Low Noise Amplifiers Exploiting Thermal Noise Cancellation*, edited by Springer.

[41] R. Kirschman, *Tutorial 8h50 lecture 1: Survey of low-temperature electronics,* WOLTE 11 (2014).

[42] J. Bardin, *Silicon-Germanium Heterojunction Bipolar Transistors for Extremely Low-Noise Applications*, Ph.D. thesis, CalTech (2009).

[43] K. Das, *Low temperature microelectronics design for digital readout of single electron transistor electrometry,* (2013).