POLITECNICO DI TORINO

MASTER's Degree in NANOTECHNOLOGIES FOR ICTs

Politecnico di Torino

MASTER's Degree Thesis

SINGLE ELECTRON TRANSISTOR ON FD SOI FOR SPIN QUBITS SENSING

Supervisors

Prof. GIANLUCA PICCININI

Prof. ADRIAN MIHAI IONESCU

Doc. FABIO BERSANO

Candidate

FRANCO DE PALMA

February - August 2021

Summary

This Master Thesis describes the design and the fabrication of a *Single Electron Transistor* (SET) on a FD SOI (*Fully Depleted Silicon-on-Insulator*) substrate.
The SET is an ultra-sensitive sensor capable of measuring single charge transitions in the environment, which makes it the best candidate for the readout of a spin qubit state. Moreover, the SET illustrated in this Thesis is based on the silicon quantum dots technology and this aspect makes easier its integration with spin qubits. Such platforms already exist and will be introduced during the dissertation, but most of them are fabricated on bulk silicon and not on SOI. Under this light, the aim of this work is to validate a robust process flow to fabricate a quantum dot on FD SOI wafers with the function of SET, the leading device for the integration of more complicated devices such as spin qubits in silicon quantum dots. This is the final goal of the *Electron Spin Qubits* project inside the *NCCR spin* network [1], a project where this Master Thesis constitutes only the very first part.

A lot of emphasis is put on the characterization of the gate oxide used to create the SiMOS structure of the quantum dot to find the best fabrication recipe for a very high-quality oxide, which is the main effective mean for the reduction of the *charge noise*, one of the most detrimental parameters in single electron electronics. The results of this study allow to obtain the best fabrication recipe.
Furthermore, another focal aspect of the dissertation is constituted by the realization of ohmic contacts on the UTB (*Ultra-Thin Body*) FD SOI wafers, from the TCAD simulations for the dopants implantation to the fabrication of the contacts and their characterization. The Thesis is entirely carried out at *Nanolab* of Prof. Adrian Mihai Ionescu [2], whereas the nanofabrication process inside the *CMi* cleanrooms [3]. Both are located at the *École Polytechnique Fédérale de Lausanne*.

[1] https://www.nccr-spin.ch/

[2] https://www.epfl.ch/labs/nanolab/

[3] https://www.epfl.ch/research/facilities/cmi/

Acknowledgements

I would like first to thank professor Adrian Ionescu, who trusted me and gave me the opportunity of joining the great Nanolab family.
Every member of the Lab helped me a lot during these months, so I have to express my gratitude to all of those people who dedicated some of their precious time to me.
First of all, the doctoral assistant Fabio Bersano, responsible for this research project and my direct supervisor. We worked together and supported each other in all the critical steps of the project and without him this work would not exist.
Then, I would like to extend my appreciation to the other members of the group, such as the post-doc Yingfen Wei, the PhDs Sadegh Kamei, Luca Capua, Carlotta Gastaldi, Felix Risch and many others.
I wish to express my gratitude also to professor Martino Poggio and Dr. Floris Braakman from the University of Basel, who continuously supported the project with their precious advice.
In conclusion, I would also like to show my deep appreciation to my friends and all my family, who have always been closed to me in the most difficult and challenging circumstances.

Table of Contents

List of Tables

List of Figures

Acronyms

2DEG

2-dimensional electron gas

BHF

Buffered Hydrofluoric Acid

BOE

Buffered Oxide Etch

BOX

Buried Oxide

CB

Coulomb Blockade

CMi

Centre of MicroNanotechnology

CMOS

Complementary Metal Oxide Semiconductor

CVD

Chemical Vapour Deposition

CVU

Capacitance Voltage Unit

DCE

Dichloroethene

DUT

Device Under Test

EBL

Electron Beam Lithography

ESR

Electron Spin Resonance

EDSR

Electric Dipole Spin Resonance

FD SOI

Fully Depleted Silicon-on-Insulator

FG

Forming Gases

HF

Hydrofluoric Acid

IPA

Isopropanol

MMA

Methyl Methacrylate

MOS

Metal Oxide Semiconductor

MOScap

Metal Oxide Semiconductor Capacitor

MOSFET

Metal Oxide Semiconductor Field Effect Transistor

PMMA

Poly Methyl Methacrylate

PR
Photoresist

PSB
Pauli Spin Blockade

PVD
Physical Vapour Deposition

QD
Quantum Dot

QEC
Quantum error correction

RTA
Rapid Thermal Annealing

RTP
Rapid Thermal Processing

SEM
Scanning Electron Microscope

SET
Single Electron Transistor

SMU
Source Measure Unit

SOI
Silicon-on-Insulator

TMA
Tri-Methyl Aluminum

TMAH
Tetramethylammonium Hydroxide

UTB FD SOI
Ultra-Thin Body Fully Depleted Silicon-on-Insulator

Chapter 1

Introduction

Quantum computing is one of the hottest research fields today. A lot of effort is put by research groups all around the world to find valid alternatives to classical computers, based on a technology which cannot be further improved for a long time to come. State-of-the-art classical micro-processors (like Apple A14 Bionic, produced by TSMC [1]) are based on a 5-nm technology, which means that in the MOS conductive channel only about 25 silicon atoms are present. IBM has just announced the first 2-nm chip based on nanosheet transistors [2],[3], but it is clear that the end of Moore's Law is approaching [4].

In this scenario where the size of the transistors cannot be shrunk down anymore to collocate always more of them on a chip with the aim of further increasing the computational power, new approaches are explored now-a-days.
One of them consists in quantum computation. The name is due to the fact that the information is encoded onto the wave function of a quantum-mechanical two-level system [5]. These fundamental blocks are called *quantum bits* or *qubits*. Quantum computers manipulate the information under the form of qubits to perform computations and, taking advantage of some main principles of quantum mechanics like *superposition* and *entanglement* [6], they are potentially able to solve a large set of non-classical computational problems [7].

Since different physical systems behave like two-level systems, many kind of platforms can be used for quantum computing. The most suitable ones for a large-scale integration of qubits are superconducting qubits, spin qubits in quantum dots and topological qubits based on Majorana fermions [8]. Other candidates to host qubits are trapped ions, spin qubits in nitrogen vacancies (NV) in diamond, neutral atoms in an optical lattice, waveguide quantum circuits and others [5], but for the moment companies like Intel, Google, Imec, STMicroelectronics and CEA-Leti have put more effort on superconducting and spin qubits in quantum dots, being them more

promising from the integration point of view.

The SET presented also in this Thesis is one of the means used in the silicon quantum dots technology to readout the state of the spin qubits and it is realized on a SOI substrate to investigate whether the SOI technology can represent an improvement with respect to bulk silicon also in the quantum technology field, like it already is for standard CMOS devices and not only.

Under this light, the aim of this chapter is to give a quick overview of the fundamentals of quantum mechanics and spin qubits to introduce the reader to the world of quantum computing, a field that, even if outside the purpose of this Thesis, represents the main application of the SET and so deserves few introductory words. In conclusion, both the working principle of the SET, realized with the quantum dot technology, and the needing of a full gate oxide characterization are introduced here and then described in details in Chapter 2.

1.1 Fundamentals of Quantum Mechanics and Quantum Computation

Quantum Mechanics is one of the youngest theories in Physics. Born at the beginning of the XX century to explain phenomena which could not be consistently described by classical mechanics such as the black-body radiation and the photoelectric effect [9], quantum mechanics is based on the principle that physical quantities can assume only *discrete* values (multiple integers of a *quantum*, i.e. their minimum assumable quantity) and it is actually the most reliable theory to describe Nature at atomic and sub-atomic scale.

Other founding principles are the *wave-particle duality*, *superposition* and *entanglement*. The first one states that quantum physical entities (like light or electrons) both behave like particles and waves [10], which in the context of this Thesis means that electrons in a quantum dot, even though having a particle nature, can be described through a wave function whose square modulus gives the probability of finding them in a certain place in the space. This last aspect is the so called *"Copenhagen Interpretation of Quantum Mechanics"* by Niels Bohr, Werner Heisenberg and Max Born [11].

Superposition means that a quantum state can be built from the linear combination of other distinct quantum states [6], or, in other words, that a quantum physical quantity can be described by a state that can assume more than one value at the same time, until it is measured (this is known as *wave function collapse*).

And the third one, the entanglement, takes place when the quantum state of a particle cannot be described independently of the quantum state of the other particle(s) of the system. The quantum state of the system as a whole is in a definite state, but the parts of the system are not [12]. In other words, two particles are entangled if there is a so strong connection between them that the result of the measurement on one affects the measurement on the other one, even if the two particles are put at a very long distance before the measurements. This means that if there is, for example, a system of two electrons where the total spin is zero, in the moment in which the spin of the first one is measured ($\frac{1}{2}$ or $-\frac{1}{2}$), the spin of the second becomes immediately known ($-\frac{1}{2}$ or $\frac{1}{2}$), no matter how far the two electrons are.

These last two properties are the ones which make quantum computation able to solve problems whose complexity grows exponentially with the system size. And this is a key aspect because systems which such properties can be used to physically implement algorithms to solve problems like optimal search for specific entries in an unordered dataset (Grover's algorithm [13]), prime factorisation of integers in near polynomial time (Shor's algorithm [14]), logistic-based problems derived from the famous *travelling salesman* problem etc [5].

Moreover, since Nature behaves quantum-mechanically, the best way to properly simulate its behaviour is through a computer which makes computation on a quantum-mechanical basis (i.e. a *quantum computer*), as already predicted by Richard Feynman [15].

DiVincenzo Criteria

Not every two-level quantum-mechanical system can be used to perform large-scale quantum computation. Some criteria, developed by David DiVincenzo in 2000 [16], must be fulfilled. They can be summarized as the following:

1. The system which hosts the qubits has to be a *scalable physical system with well characterized qubits*, where "characterized" means that its Hamiltoninan (and so all the internal energy levels and the coupling between them) should be accurately known, as well as the interactions with other qubits and external fields.

2. The qubits must have the ability to be *initialized to a simple fiducial state*. In fact, also in classical computation, registers have to be initialized to a known value before starting the computation. In the case of qubits it is the same, where initialization usually consists in having all the qubits in the ground state through a *cooling* operation. The initialization to the ground state is also

a key aspect to apply all the algorithms of *quantum error correction* (QEC) [17].

3. Qubits with *long relevant decoherence times, much longer than the gate operation time*. Since quantum systems are very sensitive to the environment, the information they carry is subjected to decoherence in time. This decoherence time has to be larger than the time needed for a quantum gate to perform an operation. How much this time has to be larger is imposed by QEC codes, algorithms which use other *ancilla* qubits, initialized to the ground state, to correct errors in computation due to decoherence. Since these ancilla qubits have to be continuously refreshed, this time is found to be $10^4 - 10^5$ times the "clock time" of the quantum computer.

4. *A "universal" set of quantum gates* constituted by a very small set of 1- and 2-qubit gates. The requirements is that it has to be possible to build a complete set of quantum gates using the internal interaction between them.

5. The system must have *a qubit-specific measurement capability*, which in other words means that an ideal measurement of a qubit state should not be affected by the nearby qubits and should keep the rest of the qubits unchanged. Moreover, if the system allows for non-destructive projective measurements, then, in principle, this can be used for state preparation of criterion 2.

All these criteria are satisfied by electron spin qubits in silicon quantum dots, the platform where the SET of this Thesis is used as readout instrument.

1.2 Electron Spin Qubits in Quantum Dots

In spin qubits the quantum two-level system is represented by the intrinsic angular momentum of an electron, i.e. its spin. The overall state of this system can be described by the linear combination of the two quantum states spin-up and spin-down.

Presented the first time by Loss and DiVincenzo [18], in such a platform electrons are confined in *quantum dots* (QD), 3-dimensional "boxes" able to trap electrons inside through potential barriers which can be tuned by lithographically defined electric gates. These systems behave like artificial atoms, because electrons can occupy only well defined (and discrete) energetic levels, whose space between them depends on the effective mass of the particle and the dimension of the dot, usually from 10 to 100 nm [8].

Thanks to the small dimensions of the dot, in principle single-electron occupancy can be achieved and, to obtain the computational basis spin-up/spin-down, an external

magnetic field is applied, so that energetic levels are split in two sub-levels occupied each one by one spin-up electron or spin-down one, separated by the *Zeeman* energy.

This is the simplest realization of a spin qubit (single-spin qubit, i.e. one electron in one dot), which can be obtained also manipulating the spin of an excess electron in an already populated quantum dot. However, also multi-dot implementations exist: the most diffused one is the singlet-triplet spin qubit, which consists in coupling two electrons in two dots by tuning the potential barrier between them.
Coupling between electrons in different dots is mediated by the electron-electron Heisenberg interaction [18], the key element for the realization of two-qubit quantum gates like CNOT, CZ, SWAP, etc.[8], which combined to single-qubit gates provide all the operations necessary for universal quantum computation.

On the other hand, single-qubit operations, i.e. spin-flip operations, are exploited by applying either a time-varying magnetic field perpendicular to the static field which splits the spin-up and spin-down energetic levels or a spatially varying magnetic field combined with a time-varying electric field. The first phenomenon is called *electron spin resonance* (ESR) and it is actuated by a microwave transmission line, whereas the second one is the *electric dipole spin resonance* (EDSR), where the bulky transmission line is substituted by integrated micromagnets and a metallic gate to apply the RF electrical signal [5],[8].

1.3 Single Electron Transistor for readout of the Qubit State

The electron spin is a magnetic moment, but very small and so difficult to sense. For this reason, two sensing methods based on a nearby charge sensor are mostly diffused. The first one is based on the *quantum point contact* (QPC) [19], but the one explored also in this Thesis consists in a QD in the proximity of the qubits array.

This QD is a conductive island usually populated by tens of electrons and isolated from the electron reservoirs connected to source and drain contacts by tunneling barriers. These barriers are so that only if the chemical potential of the dot lies within the source and drain window, an electron can tunnel from the source to the dot and then from the dot to the drain, otherwise no current flows through the device. This phenomenon is known as *Coulomb blockade* (CB), because the passage of electrons is allowed only in particular conditions and, since the control is ultra-precise (electron by electron), the result is a SET.

The SET can be capacitively coupled to the array of qubits simply by putting it

nearby. Any electrostatic change in the environment (like for example an electron which tunnels in or out from a QD of the array), shifts the chemical potential of the SET and a change in the SET current is detected.

The readout consists in putting the system in the conditions so that this electrostatic change can take place if the qubit is in a certain state (usually the excited state) or not if it is in the other (usually the ground state). The presence or absence of this electrostatic change, which directly indicates the state of the qubit to readout, is detected by the SET. The protocols to induce this change are two: the *Elzerman* protocol and the *Pauli Spin Blockade* (PSB), which will be discussed in details later in Chapter 2.2.

1.4 High-Quality Gate Oxide for "Charge Noise" Reduction

Quantum systems are very sensitive to environmental noise, which can cause detrimental effects on quantum information encoding. These effects are *relaxation*, from the excited state to the ground state, and *decoherence*, which corresponds to a randomization of the phase of the qubit state [8].
One of the causes of this noise has been identified in the traps both in the oxide and at the semiconductor-oxide interface [5]. Quantum dots are defined by applying voltages through gate electrodes and also the detuning between adjacent dots is a function of the applied voltage. These potentials can be influenced by traps, which induce fluctuations translated into a *1/f* background noise, called *charge noise*.

An effective mean in reducing this noise has been individuated in removing the thick aluminum oxide (used as blanket oxide to reduce leakages from the long interconnections) in the region of the quantum dots, leaving just the gate silicon oxide between the silicon substrate and the metallic gates [20]. A reduction was also noticed for thinner gate oxides with respect to thicker ones [21].
In general, the idea is to reduce oxide and interface traps as much as possible to the ideal value of 10^{10} cm^{-2} [22]. This value can be reached in a thermal grown oxide through a dry oxidation in *dicholoethene* (DCE) and a low temperature *rapid thermal annealing* (RTA) in *forming gases* (FG), as will be deeply discussed in a dedicated chapter. However, some fabricating process necessary for the realization of the quantum dots like *electron beam lithography* (EBL) and high-temperature RTA for implanted dopants activation can create traps in the oxide and for this reason their use should be very limited or, ideally, substituted by other less destructive fabrication techniques.

1.5 Thesis Outline

The work carried out in this Thesis covers all the aspects of the SET realization: design, simulation, fabrication and characterization. The presence of all these components suggests a division in chapters to address each one separately with a chronological order. However, the flow followed is never linear and a continuous transition between these four domains is necessary. For example, design and simulations are not necessarily preliminary to device fabrication and characterization, but some fabrication tests or characterization results can influence design and simulations choices. This will be a recurring theme during the Thesis. Under this light, the presented work is structured in the following way:

- **Chapter 2** provides the reader the theoretical background necessary to understand all the choices and results illustrated along the way. However, being the field explored by this Thesis very huge and interdisciplinary, sometimes there is not enough space to go into details. When this happens, proper references are given for particularly interested readers.
- **Chapter 3** illustrates all the process flow for the realization of the SET inside the academic cleanrooms of EPFL. Moreover, preliminary fabrication tests are already presented inside this chapter to understand some choices made in the next chapters.
- **Chapter 4** presents the complete layout and design of the SET, layer by layer and from the device level to the wafer one. Then, the results of TCAD simulations for the source and drain ohmic regions doping are illustrated.
- In **Chapter 5**, the fabrication results are reported, from ellipsometer measurements for thicknesses check and SEM pictures of the different layers of the device.
- **Chapter 6**, finally, reports all the results on the complete study of the gate oxide quality, as well as the ones concerning the realization of low resistance ohmic contacts on UTB FD SOI wafers.
- **Chapter 7** summarizes the obtained results and gives an outlook on the future plans.
- In conclusion, **Appendix A** and **B** provide further details on the MOS and the metal-semiconductor contact theory, as well as their parameters extraction.

Chapter 2

Theoretical Background

The purpose of this chapter is first to provide all the theoretical background needed to understand how QDs can be defined in the Silicon-MOS platform and how to use them to realize both a SET and also an array of spin qubits, complicating a little the design and the routing. Then, a full description of the SET is given: from *Coulomb blockade* and *charge noise* phenomena and characterization to the readout protocols for spin qubits. Moreover, the complete picture of oxide traps is provided in order to understand: the possible *charge noise* origin, the extraction of their value from C-V curves of MOS capacitors and the possible reduction (or increasing) of their value during fabrication processes. In conclusion, an overview of the SOI technology and its advantages is given.

2.1 Quantum Dots in the Silicon-MOS platform

Quantum dots in silicon are defined under the Si/SiO_2 interface thanks to the voltages applied through lithographically patterned metallic gates. For this reason, this platform is referred as Silicon-MOS or simply SiMOS. "MOS" stands for *"Metal-Oxide-Semiconductor"* and it is the structure on which is based all the CMOS and MOSFET technology, thanks to the powerful capability of this platform of making a region conductive or not, according to the voltage applied on the gate of the MOS capacitor. In this section, first a general description of confinement structures is given, then the focus is moved to QDs and their state-of-the art realization for SETs and spin qubits on the Si-MOS platform.

2.1.1 2D-Electron Gas and Quantum Wells

In crystalline solids, the periodicity of the lattice results in a periodic potential seen by the electrons. Their motion inside such materials is described by the

solution of the Schrödinger equation for a periodic potential, which consists in plane waves modulated by *Bloch* functions [23]. The periodic potential creates allowed and forbidden energy bands, but inside the former electrons are free to move in every direction of the space. This is possible because, in a perfect periodic and infinite lattice, there are no limitations on the values that the electron wavevector can assume. However, electrons can be confined in '2D' structures like heterojunctions and quantum wells (where the confinement extends in one direction), in '1D' structures like nanowires (confinement in two directions) and '0D' structures (confinement in all the three directions), i.e. QDs.
Confinement means that in that direction(s), electrons are not free to move, but their energy is quantized, i.e. discrete, because of any phenomenon able to break the lattice perfect periodicity. For example, in the simplest case, this can be a finite solid, where the boundary conditions force the electron wavevector to assume discrete values. This is exactly what happens in the real world, where solids of course have finite dimension, but still too large to see this quantization effect. A more concrete case is constituted by putting in contact two different materials. At the interface, the Fermi levels in the two material have to match through an adjust in the net charge density. This latter forces the energy bands to bend and create an electric field which can confine the electrons in a thin sheet of charge close to the interface.

Heterostructures

The most simple case of bands misalignment is obtained in heterostructures, where a lower bandgap material (like GaAs) is put in contact with an higher bandgap material (like AlGaAs) [24]. AlGaAs is n-doped, so that electrons from the donor

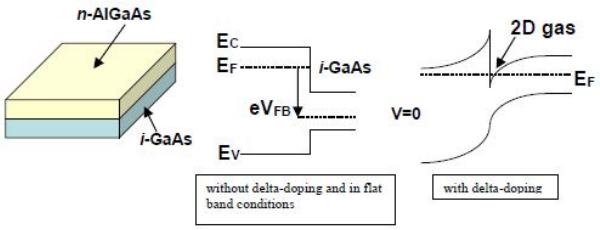


Figure 2.1: GaAs-AlGaAs heterostructure: band structure before contact and at equilibrium [24].

states, instead of being promoted to AlGaAs conduction band by thermal energy, prefer leaving the material and move directly to the conduction band of intrinsic GaAs, being at lower energies. In this way, the donor atoms in the AlGaAs remain positively charged and create an electrical field (and a band bending) which confines

the electrons in GaAs in a quantum well close to the surface, as shown in Fig.2.1. This phenomenon is known as *delta doping* or *remote doping*, since the material of interest, GaAs in this case, is electron-doped, but the impurities introduced for this purpose are all in the wider bandgap material, because GaAs is intrinsic. The result is a 2D-electron gas, with a large electron mobility, where electrons are free to move in both the directions parallel to the surface, but not in the perpendicular one. In this vertical direction, the energy of the electrons is quantized and their wavefunction is described by *Airy functions* [25]. As a consequence of this, electrons lie in what are called *sub-bands*, because their energy is continuous inside each sub-band, but there are more of them at different discrete energy in the confinement (vertical with respect to the surface) direction.

SiMOS Capacitor

A similar effect can be obtained at the silicon-silicon oxide interface in a MOS structure. Also in this case, the difference between the Fermi levels in the metal and in the semiconductor induces a band bending to re-establish the equilibrium, as shown in Fig.2.2 [22]. In this structure, what happens is that if the silicon is for

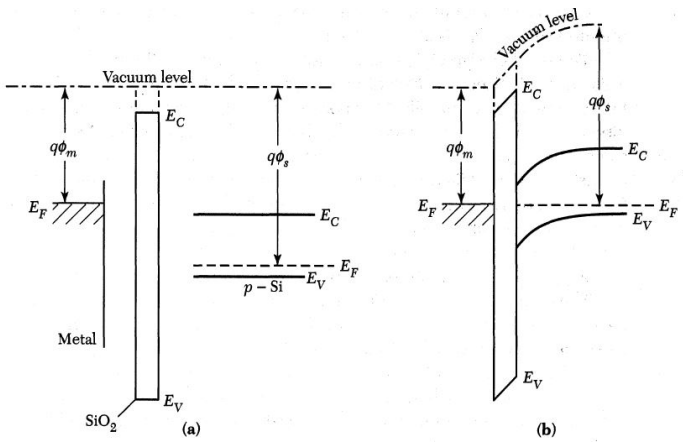


Figure 2.2: p-type Silicon-MOS capacitor: band structure before contact (a) and at equilibrium (b) [22].

example a p-type one, i.e. where the substrate is doped with acceptor impurities to increase the concentration of holes in the valence band, when negative gate voltages are applied, first bands from downward bent become flat (*flatband condition*), then, for higher negative voltage values, they start bending upward and holes are accumulated at the Si/SiO_2 interface (*accumulation* regime). On the other side, for positive gate voltages, downward band bending is enhanced: first holes start depleting from the interface (*depletion* regime) and then, when the voltage drop

on the semiconductor (known as *surface potential* ψ_S) exceeds the bulk potential ψ_B, i.e. the difference between the Fermi level and the intrinsic Fermi level in the substrate, the carrier population is inverted between majority (holes) and minority (electrons) carriers. This is the *inversion* regime. And by further increasing the voltage, when the semiconductor is fully depleted of majority carriers, ψ_S saturates to $2\psi_B$ and the minority carriers concentration at the surface equal the majority ones in the bulk (value which usually corresponds to the substrate doping). This last regime is known as *strong inversion* and the voltage to apply to have this condition is called *threshold voltage* and it is given by:

$$V_{th} = V_{FB} + \frac{\sqrt{2\varepsilon_s q N_A (2\psi_B)}}{C_{ox}} + 2\psi_B, \tag{2.1}$$

where V_{FB} is the *flatband potential* and the *bulk potential* ψ_B is given by the Shockley equation:

$$\psi_B = \frac{k_B T}{q} ln(\frac{N_A}{n_i}). \tag{2.2}$$

For a n-type MOS capacitor (n-substrate) the principle is exactly the same, but the voltages to apply have opposite signs to establish the same operation regimes, i.e. to create the inversion layer, made of holes, negative voltages have to be applied (V_{th} is negative).

Fig.2.3 shows the p-type MOS band diagram in the inversion regime: when the

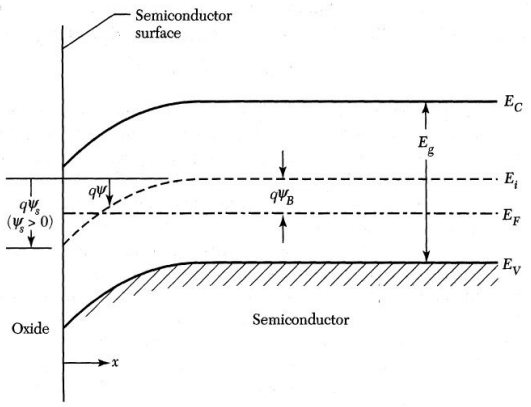


Figure 2.3: p-type Silicon-MOS capacitor: band structure in inversion (positive gate voltage applied) [22].

Fermi level passes below the intrinsic one, the carrier population starts inverting. Further bending makes the conduction band very close to the Fermi level, the last occupied energetic level. The result is, also in this case like in the previous heterostructure example, a 2D-electron gas confined at the surface. The difference

with respect to the previous example is that this sheet of charge is very thin (the electron concentration drops from 10^{13} to 10^{11} cm^{-2} in few nanometers [26]), voltage induced and with a lower mobility because of the impurities at the Si/SiO_2 interface.

Debye Length

Before understanding how to extend confinement also the other two direction to get a QD, it is fundamental to understand which are the physical dimensions for which confinement and quantization start playing a role. The basic principle is that, if a system confined in one direction is considered, the energy levels in that direction have to be separated of an amount which is larger than the thermal energy. If this requirement is not fulfilled, the separation between the different energetic levels disappears and energy becomes a continuous quantity again also in the confinement direction.
The computation that is usually done consists in making the thermal energy smaller than the energy of the first energy level of a quantum well [24]:

$$k_B T < \frac{\hbar^2 \pi^2}{2 m_{eff} L^2}, \tag{2.3}$$

where m_{eff} is the effective mass of the electron in the material, $\hbar$ the Planck constant, k_B the Boltzmann constant and L is the dimension of the well. The result is:

$$L < \lambda_{DB} = \frac{h}{\sqrt{8 m_{eff} k_B T}}, \tag{2.4}$$

where λ_{DB} is known as *Debye length*. To have confinement, the dimension of the well has to be smaller than the Debye length, which increases by decreasing the operating temperature. If the case of silicon is considered, it can be noticed that at room temperature confinement would manifest only for dimensions below few nm. This means that in conventional CMOS devices, even if the 2DEG is closed to the Si/SiO_2 interface, confinement never happens and it starts playing a role only in the very last technological nodes. On the other hand, in materials like GaAs and InAs confinement begins for larger dimensions with respect to silicon at the same temperature. This is due to the fact that electrons in such materials have a lower effective mass (and so, also an higher mobility, as already introduced before). However, if cooled down to 4.2 K (the boiling point of He^4), also in silicon quantum confinement starts manifesting in wells smaller than about 60-80 nm. For this reason, QDs have a typical size of 50 nm or less.

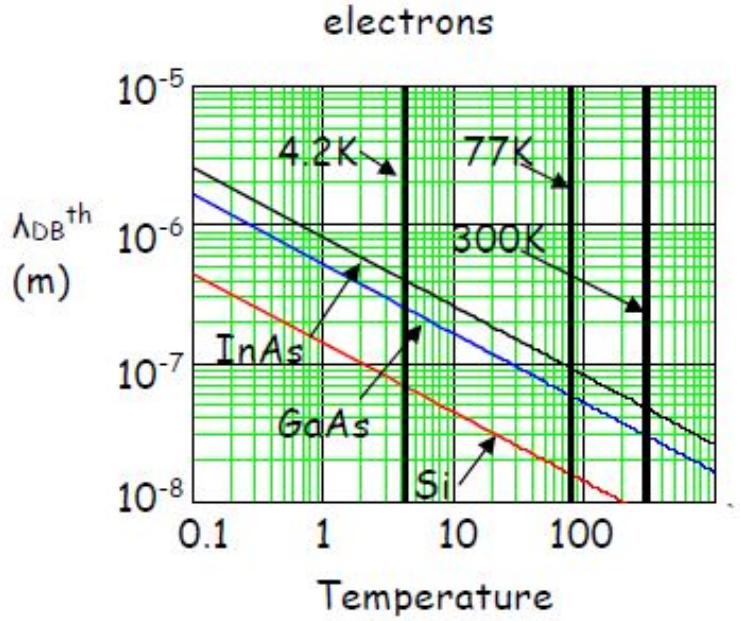
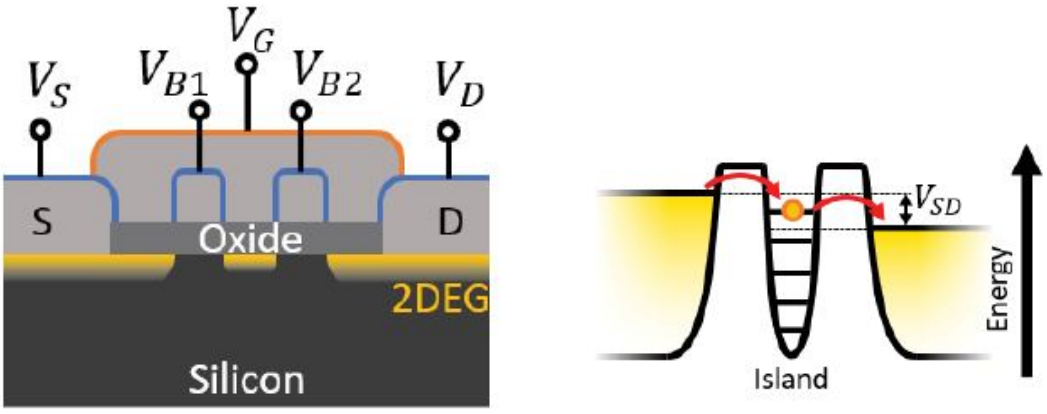


Figure 2.4: Debye length vs temperature for different semiconductor materials [24].

2.1.2 Quantum Dots

After understanding how electrons can be confined in a 2D structure, the next step for the realization of a QD is to extend this confinement also in the other two remaining directions. This further confinement is known as *lateral confinement* and it is imposed by lithographically defined metallic gates. The most basic example of a QD in silicon defined in such a way is illustrated in Fig.2.5a [5]. A *top gate* V_G accumulates electrons at the semiconductor-oxide interface as explained in the previous section. Then, two *barrier gates* V_{B1} and V_{B2} locally deplete the channel from electrons, creating two tunneling barriers whose tunneling rate can be tuned through the applied voltage. In this way, a QD is created in the 2D-electron gas between the two tunneling barriers: vertical confinement is imposed by V_G which confines electrons at the interface, whereas lateral confinement is imposed in one direction by the two barrier gates and in the other direction by the width of the top gate V_G, which is less than 50 nm. Typical dimensions of the dot are usually 50 nm x 30 nm (in the longitudinal direction) x few nm (in the transversal direction, the one of the 2DEG). The first dimension, in the source-drain direction, is usually computed as the distance between the centre of the two barrier gates, usually 30 nm wide and 20 nm far, and so the result is about 50 nm. The second one, always in plane, but in the perpendicular direction with respect to the previous one, is



(a) **(b)** *b*

Figure 2.5: SiMOS cross section of a quantum dot and schematic of its chemical potential. a) V_G creates the 2DEG at the Si/SiO_2 interface, V_{B1} and V_{B2} locally deplete the channel to define the QD. The source and drain ohmic contacts V_S and V_D on the doped regions complete the SET. b) Schematic of the chemical potential of the SET: if one of the energetic quantized levels of the dot lies within the source and drain window, an electron can tunnel from S to D and a current is measured [5].

given by the width of the top gate V_G, which is usually about 30 nm. In conclusion, the transversal dimension is only few nm, because it is the typical thickness of the 2DEG induced in a SiMOS capacitor. This last aspect is not essential in the realization of a SET, but it is when using the dot to host a spin qubit because, if the longitudinal and transversal confinement differ from an order of magnitude, the longitudinal and transversal energetic levels are well separated and do not start mixing, so that it is easier to know in which energetic level the electron is.

Due to the small dimensions of the dot, as just explained, the electron energetic levels are quantized and well defined inside the dot: for this reason, adding an electron from an external electron reservoir to the dot requires an energy E_A, called *addition energy*, equal to:

$$E_A = E_C + \Delta E, \tag{2.5}$$

where ΔE is the energy spacing between two consecutive discrete quantum levels and E_C is the *charging energy*, an energy which has to be spent in order to overcome the coulomb repulsion between the new extra electron and the ones already present

inside the dot. This value is equal to:

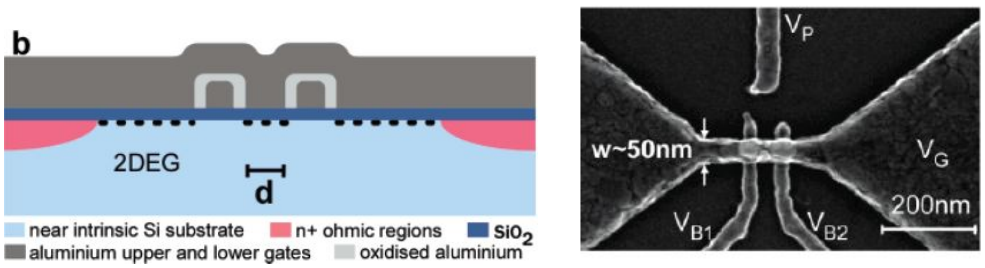
$$E_C = \frac{e^2}{2C_\Sigma}, \tag{2.6}$$

where e is the electron elementary charge, C_Σ is the total capacitance of the dot. Charging energy is a peculiarity of QDs and its physical consequences, being the basis of all QDs technology and applications, are explained in detail in Chapter 2.2.

QDs as SETs

If the structure presented in the last section is completed with source and drain implanted regions and contacts at the two extremities of the channel induced by V_G, the result is a SET (refer again to Fig.2.5a). The design is similar to the one of a MOSFET whose channel is locally depleted to define a QD, called also *island*, through the realization of tunable tunneling barriers. As shown if Fig.2.5b, when one the energetic levels of the dot is within the source and drain window, an electron can tunnel from the source to the island and then to the drain and a current is measured, as it will be described in Section 2.2, the one focused on the working principles of the SET. The energy levels in the island can be lifted up or lowered by the top gate, whereas the height of the tunneling barriers can be tuned by the barrier gates.

One of the first implementation of a SET with QDs was published by Dzurak's group in Sydney in 2007 [27]. Fig.2.6a shows the schematic of the SET, very



(a) **(b)**

Figure 2.6: Single QD used as SET, Dzurak's group implementation. a) Schematic of the cross section. b) SEM image, top view [27].

similar to the one in the previous Fig.2.5, where the metallic gates are realized in aluminum and the interlayer insulation is provided through aluminum oxide. Fig.2.6b, instead, is its *Scanning Electron Microscope* (SEM) image, where the top gate V_G and the two barrier gates V_{B1} and V_{B2} are clearly visible. V_P is a *plunger*

gate which is not used in this SET realization, but to accumulate electrons to study the SET sensitivity to a change in the electrostatic environment. The source and drain contacts are not shown here, but are realized above the doped regions, after having etched away the oxide layer.

Another example of SET realized with this design is shown in Fig.2.7 [28]. The

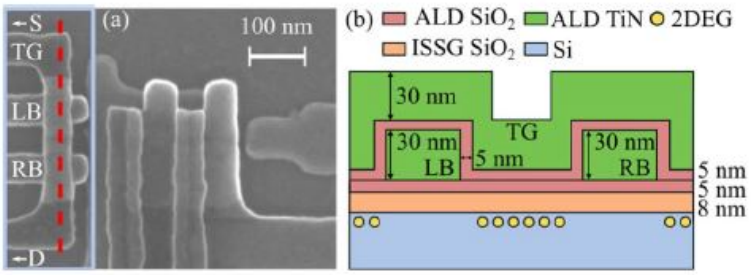
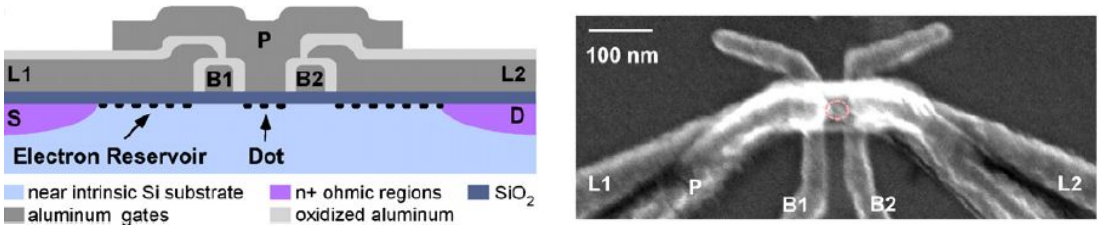


Figure 2.7: SEM image (a) and cross section schematic (b) of the SET realized by IMEC [28].

architecture is exactly the same, the only differences are in the choise of the materials: the metallic gates are realized in TiN and the interlayer oxide is silicon oxide.

An alternative possible implementation of the QD-based SET is reported in



(a) **(b)**

Figure 2.8: Single QD used as SET, Dzurak's group updated implementation. a) Schematic of the cross section. b) SEM image, top view [29].

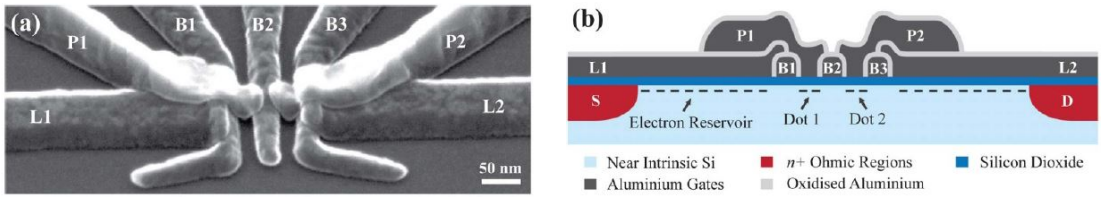
[29] and in Fig.2.8. Proposed again by Dzurak's group, it consists in splitting the role of the top gate into three separate metallic gates: the first two, called *lead gates* (V_{L1} and V_{L2}, one on the left and one on the right of the dot) have the role of two electron reservoirs, supplying the 2DEG close to the dot from the doped regions, whereas the third one is called *plunger gate* (V_P) and allows a better control of the chemical potential of the dot, being right above the dot region and not on the whole channel like the previous proposed design with a unique top gate. The reason of this more complicated design (three metallic layers instead of two) is justified by

the fact that this device was realized to study the filling of the valley-orbit states in silicon, a study which needed a precise control on the energetic levels in the quantum dot (see [29] for further details).

However, since such a study is not the purpose of the SET proposed in this Thesis, the adopted design is the first proposed one (just one top gate and two barrier gates), being simpler both from the design and the fabrication point of view. In fact, the purpose of the project is to show whether the realization of a SET on a SOI substrate with the QDs technology can bring better results in terms of reliability and charge noise and so a precise control of the energetic levels of the conductive island is not needed at this stage (it will for spin qubits implementation).

QDs as Spin Qubits

QDs used as SET usually operate with a large number of electron occupancy (N=10-100 electrons), however the can be used also as spin qubits, under the condition of working in the single electron regime (N=1-4 electrons in each dot) [5]. This kind of platform is able in principle to satisfy all the DiVincenzo criteria (see Chapter 1.1), where the computational basis spin-up/spin-down is obtained applying a static magnetic field which splits the electron energetic levels in two. Single-qubit operations in silicon are possible thanks to a ESR transmission line which creates, through an AC current, a time-varying magnetic field in the perpendicular direction with respect to the one of the static field which aligns the spin, or thanks to an array of integrated nanomagnets on the top of the qubits which creates a space-varying magnetic field to take advantage of the EDSR phenomenon. This last implementation is the one presented in Delft [8] and it is very promising from the large-scale integration point of view, being the nanomagnets more scalable with respect to the ESR transmission lines.
So, in principle, a single QD depleted to the single electron regime can be used as spin qubit. However, to get the full computational power potentially offered by quantum computing, entanglement between two or more qubits has to be established. To do so, electrons in different adjacent QDs have to interact among them. From the physical point of view, it is possible thanks to exchange (or Heisenberg) interaction, which, from the device point of view, is possible by tuning the value of the potential of the tunneling barriers which define the dots.
The most simple platform to create entanglement consists in two neighboring QDs separated by a tunneling barrier to build the singlet-triplet qubit. An example of this platform is shown in Fig.2.9 [30]. QDs are formed under the two plunger gates, whereas barrier gates laterally define them and control the tunneling rate and the interaction between the electrons in the two dots. Like in the improved design of the SET explained in the previous section, the lead gates supply electrons close to

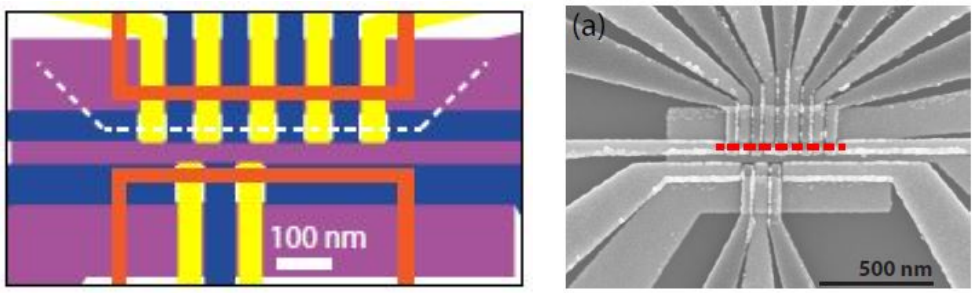


(a) **(b)**

Figure 2.9: Double QD for singlet-triplet spin qubits, proposed again by Dzurak's group. a) SEM image, top view. b) Schematic of the cross section [30].

the dots to connect them with the ohmic regions (not shown in Fig.2.10b).

When the QDs start to be more than two, the design is a little bit more complicated, i.e. more metal layers are needed to ensure proper confinement. An example is provided by the already anticipated publications from Delft and reported in Fig.2.10 [8]. In the upper part of the design, four QDs are defined under four plunger gates



(a) **(b)**

Figure 2.10: Top view of the array of four QDs and a SET on the SiMOS platform. The array is in the upper part of the figure, whereas the SET is defined in the lower part. a) Schematic of the cross section. In blue in the vertical direction are pictured the five plunger gates (one above each dot, four for the array and one for the SET); in blue in the horizontal direction the four lead gates bring the 2DEG close to the dots; in yellow are represented the seven lead gates which confine the five dots (five for the array and two for the SET); in purple the three screening gates which further laterally define the five dots; in orange the border of the two nanomagnets deposited above the whole structure for EDSR. b) SEM image of the array and the SET. The four qubits are defined under the red dashed line [8].

(the four in blue in the upper part of the schematic reported in Fig.2.10a) and confined by the five barrier gates (in yellow in the upper part). In the lower part,

a SET is defined, as in the previous already explained cases, by a plunger gate (blue) and two barrier gates (yellow). Also in this case, four lead gates (in blue, horizontally in the schematic) serve as electron reservoirs, but the new aspect is the presence of a fourth metal layer. This layer, usually deposited as very first layer, is constituted by *screening gates* (the three in violet) and they provide a better lateral confinement to the five QDs. In all the previous designs presented so far, the lateral confinement along the direction perpendicular to the current direction (from source to drain) was imposed simply by the finite a very small widths of the top/plunger gates, but when density increases this is not enough anymore. Moreover, the SET has to be only capacitively coupled to the array of QDs (see Chapter 2.2): no electrons have to tunnel from the array to the SET and vice versa. The central screening gate is exactly designed for this purpose. Other very similar realization of spin qubits in SiMOS QDs are reported in [31] and [32], whereas the same realized in Si-SiGe are reported in [21] and [20]. In Si-SiGe QDs, the confinement in the transverse direction is not imposed by a gate voltage like in SiMOS architectures, but by the tensile strain caused by the mismatch between the lattice constant of a thin Si layer compressed between two SiGe layers [21].

2.2 Single Electron Transistor as ultra sensitive charge sensor

The focus of the theoretical background moves now to the SET, described first in general, i.e. not strictly linked to the QD technology, then how it can be used to extract useful parameters to characterize QDs is explained. In conclusion, how to use it as charge sensor for spin qubits readout is presented.

2.2.1 Resonant Tunneling and Coulomb Blockade

The SET is a unique device in which a small conductive island is separated from the electron reservoirs (source and drain) by two tunneling barriers. A tunneling barrier is a potential barrier which has a non-zero probability of being crossed by an electron. In fact, from the classical point of view, particles whose kinetic energy is lower than the height of the potential barrier do not have any chance in crossing the barrier, but quantum mechanically that particles, electrons in this case, can behave also like waves (see Chapter 1.1), thus having a finite non-zero probability of passing through the barrier. And if the barrier is not just one, but two like in the case of the SET, they can define a quantum well between them with discrete allowed energetic levels. In such a double-barrier device, conduction can occur only when the energetic levels of electrons in this conductive island are located in an energy window which allows to exchange carriers with the two electron reservoirs,

otherwise no current is measured from one terminal to the other. The tunneling probability across a single potential barrier can be approximated to an exponential function which depends on both barrier's parameters like its height (in energy) and width, but also on particle's parameters like its effective mass (see Appendix B). That is the reason why only very small particles have non negligible probability of crossing a very narrow potential barrier. However, when a second tunneling barrier is present, a new phenomenon must be taken into account: the interference between transmitted and reflected waves between the two barriers (again, the waves are the electrons wavefunctions). Taking this into account, if the two barriers are identical, the tunneling probability of one electron to cross one barrier is given by [24]:

$$T = \frac{\Gamma^2}{\Gamma^2 + (E - E_R)^2}, \tag{2.7}$$

where Γ is the tunnel rate through the barriers (defined by both the barrier width and height and linked to the lifetime τ of an electron trapped between the two barriers by the Heisenberg uncertainty principle: $\Gamma \cdot \tau = \hbar$) and E_R is the *resonant energy*, a certain value of energy for which the tunneling probability is maximum even tough the tunnel rate through the barriers taken singularly is very small. This phenomenon for which a particle has a maximum probability of crossing a potential barrier if it has exactly an energy equal to the resonant energy is called *resonant tunneling*. Fig.2.11 shows the band diagrams of a resonant tunneling device out of resonance (a), in resonance (b) and in the conventional tunneling regime, i.e. when the two barriers are so far in energy that no interference takes place (c).

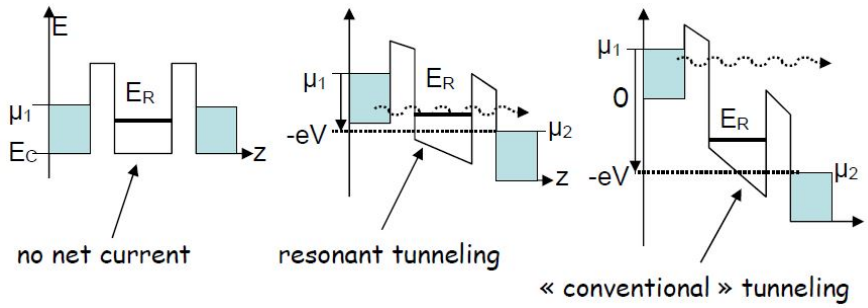


Figure 2.11: Band diagram of a resonant tunneling device out of resonance (a), in resonance (b) and in case of conventional tunneling (c) [24].

However, the above explained phenomenon of resonant tunneling is not enough to explain why the SET can sense single charges. In fact, for the moment, what is clear is just that electron tunneling is allowed only when the resonant energy E_R lies within the energy window of source and drain, but when this happens a continuous current from source to drain is originated, which can be computed by

integrating the tunneling probability in that energy window.

To fully explain the SET physics, another phenomenon has to be taken into account, knows as *Coulomb blockade*. In fact, when an electron tunnels from one reservoir to the central island or from the central island to the other reservoir, the electrostatic potential inside the island is modified. This means that when an electron tunnels into the dot, the potential energy of the electrons inside is a little raised, because of the repulsing interaction between electrons. Usually this phenomenon can be fully neglected because the discrete energetic levels inside the dot are spaced more than this electrostatic energy, but not in a SET, where the capacitance is very small due to the small size of the island. In conclusion, what happens in a SET is that the potential energy modification brought by adding just one electron is much larger than the spacing between the energetic levels, which means that the tunneling current is dictated by the electron charge. And Coulomb blockade (CB) is exactly the phenomenon for which tunneling is forbidden at certain voltages because of Coulomb repulsion between electrons.

Capacitance Model of Coulomb Blockade

To model the CB, there is no need of quantum mechanics. A simple capacitance model is sufficient, where the two tunneling barriers (the two barrier gates in the QD implementation of the SET) are modelled through "leaky" capacitors C_1 and C_2, i.e. dielectrics which allow the charge passage, whereas the action of the top gate, which controls the dot occupancy, through a conventional capacitor C_G, as pictured in Fig.2.12, where $N|e|$ is the total number of electrons inside the island.

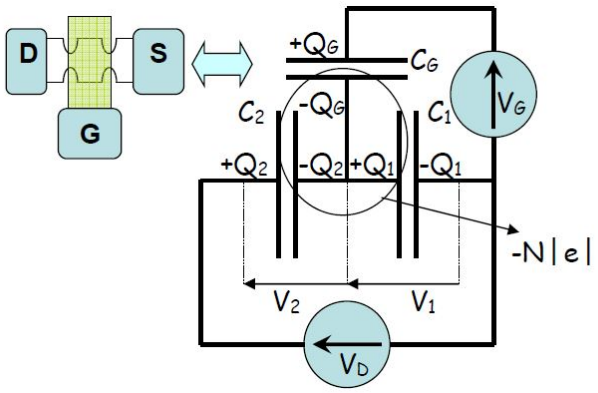


Figure 2.12: Capacitance model of the SET with a capacitor C_G between the gate and the conductive island and two capacitors C_1 and C_2 between the two source and drain reservoirs and the island [24].

The basic laws of electrostatic allow to write down all the equations to describe the circuit:

$$Q_G = C_G(V_G - V_1); Q_1 = C_1V_1; Q_2 = C_2V_2; Q_1 - Q_2 - Q_G = -N|e|. \quad (2.8)$$

The electrostatic energy of the dot is given by:

$$E = \frac{1}{2}\int \rho(\vec{r})V(\vec{r})d^3\vec{r} = \frac{1}{2}(Q_GV_G - N|e|V_1 + Q_2V_D), \quad (2.9)$$

so to compute it, the system of Eq.2.8, has to be solved to find V_1 and V_2 as functions of all the other parameters, being $V_D = V_1 + V_2$:

$$V_1 = \frac{1}{C_1 + C_2 + C_G}(C_2V_D + C_GV_G - N|e|), \quad (2.10)$$

$$V_2 = \frac{1}{C_1 + C_2 + C_G}(N|e| + (C_1 + C_G)V_D - C_GV_G). \quad (2.11)$$

Substituting these last two equations into Eq.2.9 leads to the complete expression of the energy in the dot:

$$E = \frac{1}{2(C_1 + C_2 + C_G)}(C_1C_2V_D^2 + C_1C_GV_G^2 + C_2C_G(V_G - V_D)^2 + (Ne)^2), \quad (2.12)$$

where the only term N-dependent is the *charging energy* introduced in Section 2.1.2:

$$E_C = \frac{N^2e^2}{2(C_1 + C_2 + C_G)}. \quad (2.13)$$

For example, adding one electron from the source if already N electrons are present costs:

$$\Delta E_C(N + 1) - \Delta E_C(N) = \frac{e^2}{C_1 + C_2 + C_G}(N + \frac{1}{2}). \quad (2.14)$$

A complete energetic balance has to take into account also the energy spent by the generators V_G and V_D to add this extra electron:

$$\Delta E_{gen} = \Delta E_G + \Delta E_D = -\delta Q_GV_G - \delta Q_2V_D = -C_G\delta V_1V_G - C_2\delta V_2V_D. \quad (2.15)$$

Substituting Eq.2.10 and 2.11 in this last equation and taking into account that the "δ" means that only the N-dependent terms remain when subtracting the two terms V_1 and V_2 in the $N + 1$ and N cases, gives:

$$\Delta E_{gen} = -\frac{|e|}{C_1 + C_2 + C_G}(C_GV_G + C_2V_D). \quad (2.16)$$

And so, the overall change in energy is:

$$\Delta E_{tot} = \Delta E_C + \Delta E_{gen} = \frac{|e|}{C_1 + C_2 + C_G}(|e|(N + \frac{1}{2}) - C_G V_G - C_2 V_D). \quad (2.17)$$

If this term is negative, it means that it is energetically favorable the tunneling of one electron from the source to the dot. This happens, if $V_D > 0$, when:

$$V_D > (N + \frac{1}{2})\frac{|e|}{C_2} - \frac{C_G}{C_2}V_G. \quad (2.18)$$

To find when also the tunneling of one electron from the dot to the drain is energetically favorable, the computation to do is similar, taking into account that the dot this time passes from N to N-1 electrons. The overall change in energy is given by:

$$\Delta E_{tot} = \frac{|e|}{C_1 + C_2 + C_G}(-|e|(N - \frac{1}{2}) + C_G V_G - (C_1 + C_G)V_D). \quad (2.19)$$

So, withdrawing one electron from the dot to the drain is energetically favorable ($\Delta E_{tot} < 0$) for (if $V_D > 0$):

$$V_D > -(N - \frac{1}{2})\frac{|e|}{C_1 + C_G} + \frac{C_G}{C_1 + C_G}V_G, \quad (2.20)$$

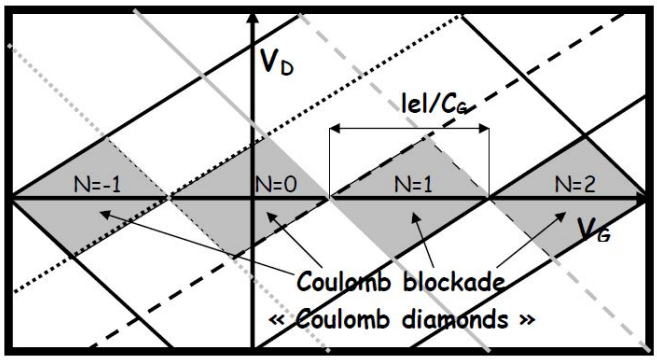
In fact, if $V_D < 0$, it is easier to inject into the dot one electron from the drain than from the source. At the same time, for $V_D < 0$, it easier to withdraw one electron from the dot to the source than to the drain. And so the two conditions, after some computations similar to ones of the two previous cases, become:

$$V_D < -(N + \frac{1}{2})\frac{|e|}{C_1 + C_G} + \frac{C_G}{C_1 + C_G}V_G, \quad (2.21)$$

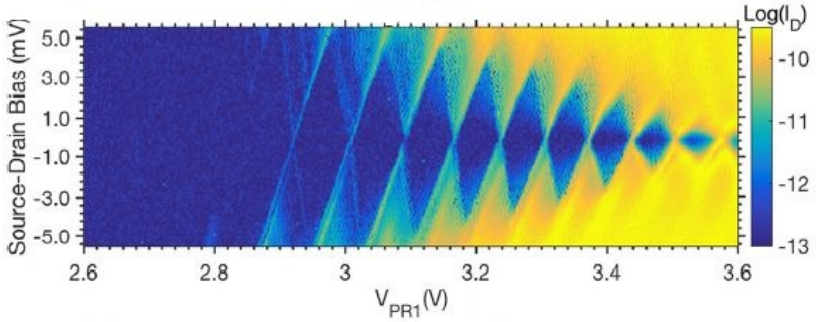
and:

$$V_D < (N + \frac{1}{2})\frac{|e|}{C_2} - \frac{C_G}{C_2}V_G. \quad (2.22)$$

These four equations are equations of lines in a V_G vs V_D graph. They define the regions in which conduction is possible and the ones in which it is forbidden. The regions in which no current passage is allowed are called *Coulomb blockade domains* or *Coulomb diamonds* and there is one of them for every N value. However, it has to be highlighted that since all the results of this discussion are derived by a simple energetic balance of electrostatic charges, N does not correspond to the absolute total number of electrons in the dot, but to the number of electrostatically uncompensated electrons. For example, the diamond corresponding to N=0 does not imply that in the dot zero electrons are present, but it means that because of



(a)

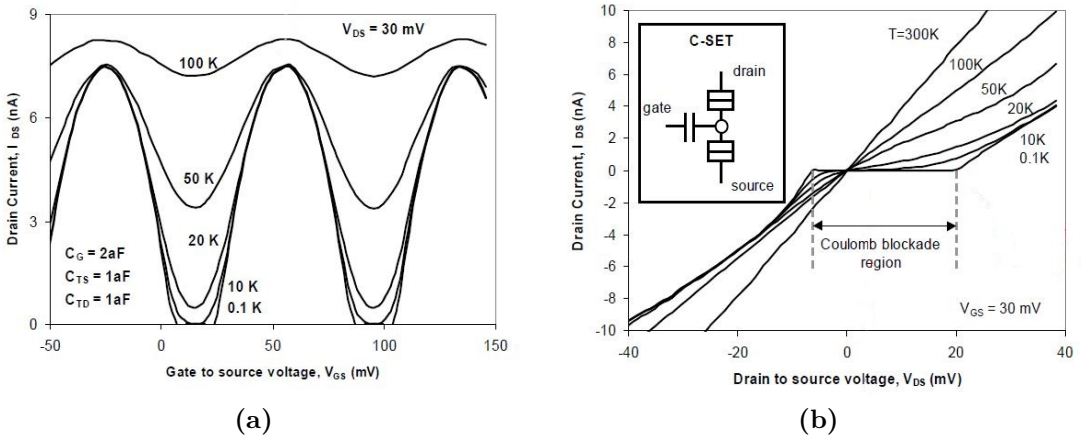


(b)

Figure 2.13: Stability diagram showing the phenomenon of CB in a SET. Inside the coulomb diamonds, it is forbidden to add or withdraw an electron if N electrons are already inside (stable regions). These diamonds show a periodicity of $\frac{|e|}{C_G}$. a) Theoretical coulomb diamonds according to the previous calculations [18]. b) Real graph of the CB showing SET current (or conductance) as a function of V_G and V_D. The disappearance of coulomb diamonds for lower gate voltages indicates that no other electrons can be withdrawn from the dot, i.e. the single electron occupancy is reached. The real diamonds also decrease their size by increasing V_G because of the increasing in the electron occupancy (and so, also the dot size) [32].

impurities which can trap or release electrons the net charge in that diamond is zero. This is the reason why, it is possible to pass from $N = 0$ electrons to $N = -1$ inside the dot, as shown in Fig.2.13a, which represents the famous *stability diagram* of a SET. Inside the diamonds, the electron passage through the dot is forbidden because of coulomb repulsion. For example, in the point corresponding to the origin ($V_G = 0$, $V_D = 0$), all the electrons inside the dot are electrically compensated ($N = 0$); however, adding (or withdrawing) an electron into (or from) the dot costs energy. Adding an electron if zero electrons are inside the dot costs $E_C = \frac{e^2}{2C_G}$,

which means that if $V_D = 0$, a gate voltage of $V_G = \frac{|e|}{2C_G}$ has to be applied. If V_D is kept to 0 V (or lower than the top of the diamonds), sweeping the gate voltage V_G has the effect of entering and going out of the different Coulomb diamonds with a periodicity of $\frac{|e|}{C_G}$. This means that in a I_D vs V_G graph peaks of current (or conductance) in correspondence of $V_G = \frac{|e|}{2C_G}$, $\frac{3|e|}{2C_G}$, $\frac{5|e|}{2C_G}$ etc. are expected, i.e. in the points were the different diamonds touch each other, as shown in Fig.2.14a [33]. Fig.2.14a highlights also another key aspect of the SET: the effect of temperature.



(a) **(b)**

Figure 2.14: Transfer characteristic (a) and output characteristic (b) of a SET. In the left graph, Coulomb blockade manifests through periodic oscillations of the drain current sweeping the gate voltage, whereas in the right one, through a region where the drain current is zero even though the drain voltage is not. For CB to be observed, cryogenic temperatures are needed for conventional dot sizes of decades of nm [33].

In fact, for CB to be observed, the thermal energy $E_T = k_B T$ has to be smaller than the minimum charging energy $E_C = \frac{e^2}{2(C_G+C_1+C_2)}$. For this to be true, cryogenic temperatures are needed because of the actual limit in fabricating very small dots (i.e. with a very small capacitance). In fact, a SET working at room temperature would require a capacitance of fractions of aF, which correspond to a dot smaller than 1 nm. Fig.2.14b shows, on the other hand, the output characteristic of the SET, i.e. the behaviour of the drain current I_D vs the drain voltage V_D. When the thermal energy does not exceed the charging energy, a clear window in which the current is zero is observed (corresponding to a vertical movement inside a coulomb diamond in the stability diagram), i.e. showing the CB. However, when the temperature provides enough energy to the system, coulomb repulsion between electrons is always overcome and the behaviour is linear like in a MOSFET at low drain voltages (ohmic or linear region).

2.2.2 Quantum Dot Bias Spectroscopy

This section concludes by illustrating how the stability graph explained above (I_D - or the drain conductance G_D - vs V_G and V_D) can be used to extract all the parameters to characterize a QD. This procedure is known as *quantum dot bias spectroscopy*. Once the stability diagram is extracted from low-frequency measurements, the gate capacitance can be extracted from the width of the coulomb diamonds, being $\Delta V_G = \frac{|e|}{C_G}$. However, this is not a so useful information, since the total capacitance of the dot is given also by the source and drain capacitances. As a consequence, what is usually done is to extract the height of the coulomb diamonds, which takes into account also the other capacitances. In fact, looking back to the equations of the lines which define the diamonds, it can be noticed that the superior vertices of the diamonds are included between $\frac{|e|}{2(C_1+C_G)}$ and $\frac{|e|}{2C_2}$, where C_1 and C_2 are usually larger than C_G. So, it is a good approximation to extract the charging energy of the dot, being $E_C = \frac{e^2}{2(C_1+C_2+C_G)}$, directly from the height of the coulomb diamonds (if the diamonds decrease their size by increasing V_G, the first one is taken) through the relationship [21]:

$$E_C = eV_{DS}, \tag{2.23}$$

being V_{DS} the voltage difference between source and drain at the superior vertex of the first diamond. Once the charging energy is known, also the dot capacitance is and from this value the size of the dot can be estimated, being the capacitance just dependent on geometrical parameters. A good approximation for a QD defined at the Si/SiO_2 interface is to treat the dot as a disk of radius R, so that: $C_{dot} = C_G + C_1 + C_2 = 8\varepsilon_{eff}R$, where ε_{eff} is the arithmetic mean of the two dielectric constants of Si and SiO_2: $\varepsilon_{eff} = \frac{1}{2}(\varepsilon_{Si} + \varepsilon_{SiO_2})\varepsilon_0$ [19]. From this computation, an estimation of the radius of the dot can be obtained, which can be compared with the one (expected) defined by the patterned gates.
Then, another important parameter is the so called *gate lever arm* α, given by [21]:

$$\alpha = \frac{eV_{DS}}{V_{add}} = |e|\frac{C_G}{C_{dot}}, \tag{2.24}$$

where V_{add} is the *addition voltage*, i.e. the width of the coulomb diamonds: $V_{add} = \frac{|e|}{C_G}$. Since C_G is smaller than C_{dot}, the addition energy ($E_{add} = eV_{add}$) is larger than the charging energy, which depends on the whole capacitance of the dot. As a consequence, α is a number < 1 and its measurement unit is eV/V. α is a useful parameter to convert the applied voltages into energies, aspect which will be very useful when dealing with *charge noise* characterization (see Section 2.2.3).

2.2.3 Spin-charge conversion

Being the working principles of the SET clear, how it can be used for the readout of the spin qubit state is now explained. The SET is capacitively coupled to the array of QDs used as platform to host spin qubits. This is done by putting it in front of the array, where the only separation is a screening gate between the QDs array and the SET to ensure no electron tunneling between the qubits and the sensor. In such a configuration, a coupling capacitance between the SET and the environment induces a shift in the SET chemical potential if the charge configuration of the environment changes. In other words, if $C_{SET,env}$ is the coupling capacitance, a change in the charge in one (or more) dots of the array induces a shift in the chemical potential of the SET given by:

$$\delta\mu = \frac{\delta q}{C_{SET,env}}. \tag{2.25}$$

This means that if the gate voltage V_G is tuned so that the SET is in resonance, i.e. in correspondence of a peak in the drain current, a shift in the chemical potential induced by a charge transition in the nearby array shifts also the resonance condition, so that, if this shift is not compensated by a δV_G, the drain current drops. Fig.2.15 shows this shift in the current peaks [5]. However, how is this change in the

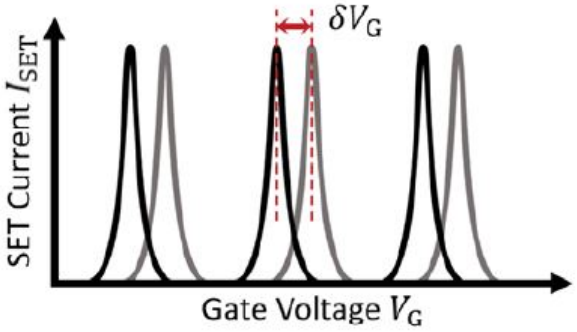


Figure 2.15: Shift in the drain current peaks due to a shift in the chemical potential induced by a charge change in the environment surrounding the SET [5].

electrostatic environment of the sensor produced? There are two main protocols based on the same principle: the system (for example a singlet-triplet qubit) is put is a condition such that if it is in a certain state, a tunneling event occur, otherwise not. This tunneling event modifies the electrostatic landscape of the system, leading to a shift in the chemical potential of the SET, so that this change is sensed. If this tunneling event does not occur, the SET does not sense any event and as a consequence the system is in the other state. These two protocols are

respectively the *Elzerman protocol* and the *Pauli spin blockade* (PSB), illustrated in the following sections.

Elzerman Protocol: Readout via an Electron Reservoir

This first approach is based on the energy separation between the spin-up and spin-down energetic levels due to an external applied electric field (Zeeman effect). In a dot depleted to the single electron regime, as in a single-spin qubit for example, if the electron is a spin-down one, it populates the lowest energetic level of the two, whereas if its spin is up, it is in the highest one. To readout the state of this electron, simply the chemical potential of the nearby electron reservoir is adjusted so to stay between these two energetic levels. More precisely, the energetic levels inside the dot are moved by the top gate, whereas the chemical potential of the reservoir is fixed. In this way, if the electron is in the spin-up state (or better, from the quantum mechanically point of view, if upon measurement of the state, the quantum system projects onto spin-up eigenstate), its energy is above the Fermi energy of the reservoir and it tunnels out from the dot. Immediately after, a spin-down electron tunnels from the reservoir to the dot, being at a lower energy. Both these two tunneling events can be detected by the SET, indicating the presence of a spin-up electron in the dot. On the other hand, if the state of the electron projects into a spin-down eigenstate, no tunneling event occurs, i.e. no changes in the SET current are detected. In both cases, the protocols end with an electron in the spin-down state inside the dot, so the operation is concluded by lifting-up the energetic levels inside the dot (always through the top gate) so that also the spin-down electron can tunnel out from the dot into the reservoir. The working principle of this protocol is shown in Fig.2.16a [8].

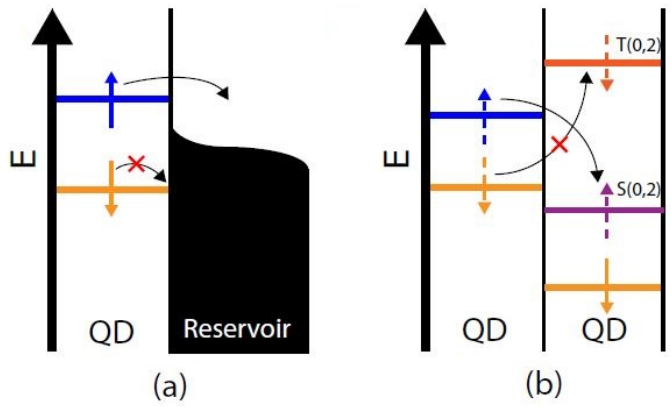


Figure 2.16: Spin state readout through Elzerman protocol (a) and Pauli spin blockade (b) [8].

Pauli Spin Blockade (PSB)

This other approach is again based on a tunneling event allowed in one condition and forbidden in another one, whose presence or absence is sensed by the nearby SET capacitively coupled to the qubits array. PSB works for one of the most common implementation of entangled spin qubits: the singlet-triplet one, where the exchange interaction between the two indistinguishable electrons adds an additional degree of freedom for the system [5]. In fact, the eigenstates of such a system constituted by two electrons in two QDs are not two, but four: one singlet state (with anti-parallel spin) and three triplet states (with parallel spin). By properly tuning the potential of the barrier gate which separates the two adjacent QDs, the coupling energy (called *tunnel coupling*) can be modified and according to this value, the lowest energy configuration of the system can be imposed: for strong positive detuning of the tunnel coupling, the system prefers to stay in a configuration where the two electrons are spread between the two dots (one electron in each QD, indicated by the notation (1,1)), whereas for strong negative detuning, the configuration with two electrons in the same QD is preferred by the system (indicated with (0,2) or (2,0)). And once the potential applied to the barrier gate sets this coupling energy, the singlet and triplet states can interchange each other in energy, as shown in Fig.2.17 [5]. In the (0,2) (or (2,0)) configuration, i.e. the

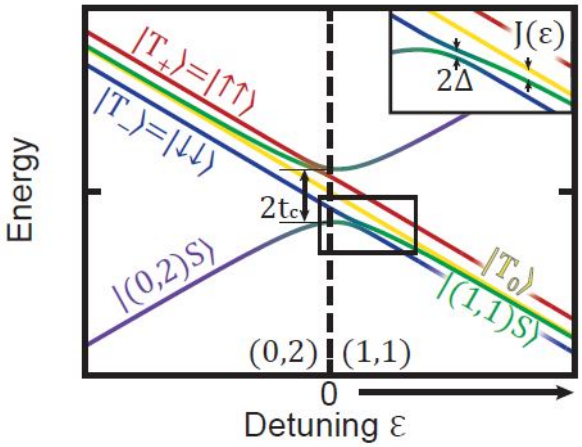


Figure 2.17: Effect of coupling energy (t_c) detuning on the energy levels of the singlet-triplet spin qubit. $J(\varepsilon)$ is the exchange term, a physical interaction which arises from the fact that electrons are indistinguishable particles, whereas Δ is the coupling term (for further details, see Ref.[5]).

one in which two electrons share the same QD, the configuration at lower energy is the singlet one, because the spin is anti-symmetric. In fact, due to the Pauli exclusion principle, two electrons cannot occupy the same energetic level if they

have the same spin. As a consequence, the triplet states (symmetric in the spin part) in this configuration are located at higher energies, because electrons with the same spin can only exist if distributed onto different energetic levels. These higher energy levels are actually excited states, which in silicon are represented by valley and orbital states (see ref. [5] for further details). On the other hand, in the (1,1) configuration, the two electrons are in different QDs, so the spin does not impose any strong constraint and all the four eigenstates are close in energy. The readout through PSB works in following way: starting from the (1,1) configuration (thanks to a positive detuning ε imposed by low external applied magnetic field and low voltage configuration), the situation is so that the triplet states are closed in energy with the singlet (1,1) state. By switching to a negative detuning, if the system was in the singlet state, an electron will tunnel into the second dot and the 2-electron system collapses into the (0,2) singlet state. If this tunneling event does not happen, this indicates that the system was in one of the (1,1) triplet states, because the (0,2) triplet states are at higher energies due to the Pauli exclusion principle explained above. In other words, the (1,1) triplet state remains *blockaded*, whereas the (1,1) singlet one can transit to the (0,2) (or (2,0)) configuration. For example, if one of the two dots, used as a reference, contains a spin-down electron, only a spin-up electron can tunnel from the other dot into the hybridized (0,2) singlet state, whereas a spin-down electron cannot tunnel to the excited (0,2) triplet state, being at higher energy. This is exactly the situation represented in Fig.2.16b. Again, as already explained, this presence or absence of electron transfer through tunneling effect is sensed by the SET, whose current will change in the first case (transfer, i.e. singlet state) and not in the second (no transfer, i.e. triplet state).

2.3 Charge Noise and its Physical Origin: Defects and Traps

Until now, only the advantages of using a SET as ultra sensitive sensor for spin qubits readout have been illustrated. In fact, the SET is a transistor for all intents, thus digital circuits can be built with an extremely low power consumption because of the nA or less drain currents [20]. Nevertheless, SET applications are for the moment limited by two aspects: the operating temperature and the noise. The first one has already been discussed, whereas the second is presented in this section, as well as its physical origin, identified in the defects and traps in the oxides and at the Si/SiO_2 interface.

2.3.1 Charge noise description and characterization

In general, it can be easily understood that, in a so sensitive device, defects, traps, ions or any kind of charge in the materials can switch the SET from being conducting or not and vice versa (this is known as *background charge effect*). As a consequence, an eventual integrated circuit constituted by plenty of SETs would require a quasi-ideal and defect free material, which is impossible for actual technologies. This problem is present also in the QDs technology illustrated so far, where the most important source of noise is represented by defects and traps in the oxides and at the Si/SiO_2 interface. These charge defects fluctuate in location over time, resulting in fluctuations in the gate voltages applied to the QDs and so in uncontrolled shifts in the chemical potential. For the SET sensitivity, it is a huge problem because these shifts in the chemical potential are read by the sensor as shifts due to qubits tunneling events, whereas they are actually due to charge defects [21]. For this reason, the background noise due to these effects is called *charge noise*, which shows a $1/f$ nature. This means that it plays a fundamental role at low frequencies, whereas at higher frequencies other kinds of noise dominate. The reason is that the traps responsible for this noise have intrinsic emission and capture times which can vary from the ms to the s range (according also the operating temperature, aspect which suggests that charge noise exhibits also a T-dependency), so that at higher frequencies the fluctuations are suppressed. This $1/f$ nature of the charge noise proves that its main origin is attributed to *f-dependent* traps, like the ones at the Si/SiO_2 interface, and C-V curves of Chapter 6 reflect exactly this aspect.

Charge noise is not only detrimental for the SET sensitivity, but also for the stability over time of spin qubits. In fact, quantum systems are very fragile and the information they carry is subjected to *relaxation* (from the excited state to the ground state) after a characteristic time T_1 and decoherence (randomization of the phase) after a characteristic time T_2^*. Several factors can influence these characteristic times, such as *Hyperfine interaction*, *spin-orbit coupling*, *phonon-induced modulation of the g-factor*, *spin-valley coupling* and, in fact, *charge noise*. The first three phenomena are the reason why QDs for quantum computation are now-a-days realized in Si or in Si-SiGe and not in GaAs anymore, even though Si, being an indirect bandgap material and with a larger effective mass (and so a lower tunnel coupling) with respect to GaAs, suffers respectively from spin-valley coupling and more from charge noise [8]. Being the first four phenomena peculiar of spin qubits and not of the SET, they are not explained here (see Ref.[8] and [5] for a complete explanation), where the focus is represented by the charge noise.

Charge noise is responsible for coupled spin qubits decoherence. In fact, its fluctuations do not directly affect the single spins, but they do when spins are coupled together through the exchange interaction, like for example in the singlet-triplet spin qubit. In fact, charge fluctuations are translated into voltage fluctuations, which affect the voltage applied by the barrier gates. And since the exchange interaction is controlled by the barrier gates, these fluctuations result in an uncontrollable exchange interaction which lead to spin dephasing [34].

Charge Noise Characterization

To characterize the charge noise, the most common approach is the so called *edge detection*. It is based on the fact that the first derivative of the drain current peaks in the V_G domain shows two maxima in correspondence of the two edges, as illustrated in Fig.2.18a [21]. In these points, little fluctuations in the gate voltage are translated into large fluctuations in current. So, first of all, after having applied a small source-drain voltage V_{DS}, 1 s time series of the drain current I_{DS} are measured for each V_G, i.e. for each value of V_G of the (I_{DS},V_G) plot, 1000 acquisitions per second of I_{DS} are taken and the average value for each time series is plotted. This operation is usually repeated for more frequencies around 1 Hz (typically 0.5, 1 and 1.5 Hz). Then, to obtain the power spectral density of the current $S_I(f)$ (measured in A^2/Hz), the Fourier transform is computed and the result of $S_I(V_G)$ at the three different frequencies is shown in Fig.2.18b [32]. This operation reported in Fig.2.18b is done to prove that the noise spectra are

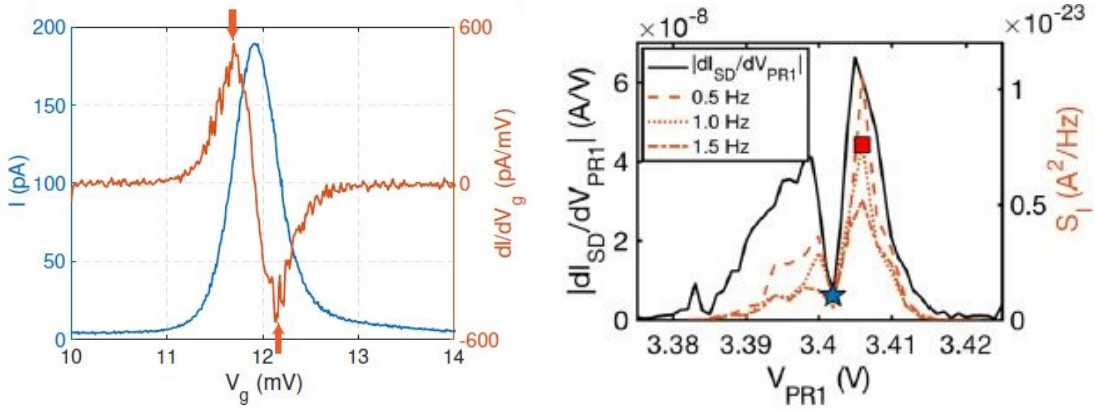


Figure 2.18: a) $I_{DS} - V_G$ curve and its first derivative: two maxima are visible on the edges, indicating the two maximum sensitivity points [21]. b) Current spectral density as a function of V_G curves at three different frequencies compared with the drain derivative curve [32].

dominated by fluctuations in the dots chemical potential and not by other noise sources (like, for example, the noise from the electronics or the environment): in fact, it can be noticed that the shape of the $|dI_{DS}/dV_G|$ vs V_G and $S_I(V_G)$ curves is the same. Then, the values of the current spectral density are taken in the point of maximum sensitivity, i.e. where the derivative is maximum (the red square in Fig.2.18b), and in the one where it is minimum (the blue star). Once V_G is fixed to these two values, the two current spectra, represented in Fig.2.19a [32], show a $1/f$ behaviour. The needing of taking two values, respectively in the maximum

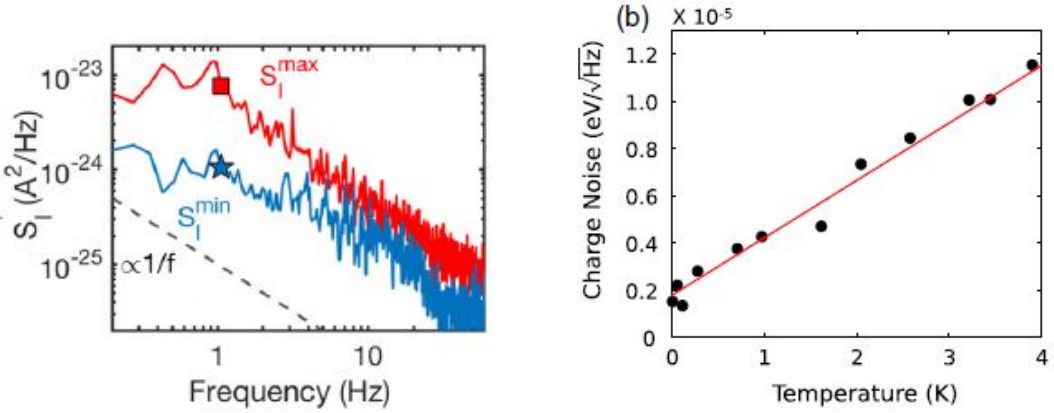


Figure 2.19: a) Current spectral density in the two points (maximum and minimum sensitivity). The frequency behaviour is a $1/f$ one [32]. b) Temperature (linear) dependence of the charge noise at low temperatures [35].

and minimum sensitivity point, is to get a differential measurement independent on the other sources of noise:

$$\Delta I(f) = \sqrt{S_{I,max}(f) - S_{I,min}(f)}. \tag{2.26}$$

However, this value of spectral noise expresses a noise in the current and not in the chemical potential. To convert it, first a current to voltage conversion is done through the slope of the derivative curve (assuming small variations, i.e. a linear dependence between voltage and current) and then the conversion into a chemical potential noise is easily obtained through the lever arm α previously extracted:

$$\Delta\mu(f) = \alpha\Delta V(f) = \alpha\left|\frac{dI_{DS}}{dV_G}\right|^{-1}\Delta I(f). \tag{2.27}$$

This value, in $eV/\sqrt{Hz}$, is the one reported in the literature (or, alternatively, its square) and it is the one referred as charge noise. However, it has to be noticed

that this is actually a noise in the potential, not really a noise in the charge. To be precise, the real charge noise, even if less used than the chemical potential noise, is obtained through the following formula, once $\Delta\mu$ and the charging energy E_c are known:

$$\Delta C(f) = (\frac{e}{E_c})\Delta\mu(f), \tag{2.28}$$

measured in $e/\sqrt{Hz}$ and usually reported at 1 Hz. In conclusion, in order to properly compare the value of charge noise extracted with the ones in the literature, also the temperature at which the measurement are taken is fundamental. In fact, the charge noise manifests a temperature dependence, thus increasing by increasing the temperature. Fig.2.19b [35] shows that charge noise at cryogenic temperatures follows a linear behaviour. Usually, even though to see the first confinement effects in Si (like for example the Coulomb blockade) a He^4 fridge is sufficient (about 4.2 K), charge noise characterization is carried out in a dilution fridge, capable of reaching temperatures down to few decades of mK. These are also the actual operating temperatures of spin qubits in QDs.

2.3.2 Oxide and Interface Traps in the SiMOS capacitor

Several times during the Thesis the physical origin of charge noise was anticipated, without, however, going into detail. In this section, the complete picture of traps and defects in SiMOS devices, individuated as the main responsible for charge noise ([8],[21],[32],[34],[35], [36]) is provided, whereas their extraction methods through C-V curves, preliminary to the study on the gate oxide carried out in Chapter 6, is explained in the dedicated Appendix A.

Traps and defects in SiMOS devices have not all the same nature. A classification of the different types exists according to their position (at the Si/SiO_2 interface or inside the oxide) and their origin, as shown in Fig.2.21 [22]. The first type of traps is represented by *interface trapped charges* at the Si/SiO_2 interface, denoted as D_{it}. These traps are constituted by interface states due to the presence of an interface between two different materials which inevitably creates dangling bonds. These interface states, in fact, have energies E_t within the Si bandgap and can be of two types: donors, if E_t is located between the Fermi energy E_F and the conduction band or acceptors, if E_t is below the Fermi energy. As a consequence, these traps can be both positively or negatively charged and are characterized by a mean capture and emission time ($\langle\tau_c\rangle$ and $\langle\tau_e\rangle$ respectively) which follows the statistics [34]:

$$\frac{\langle\tau_c\rangle}{\langle\tau_e\rangle} = e^{(E_t - E_F)/k_B T}. \tag{2.29}$$

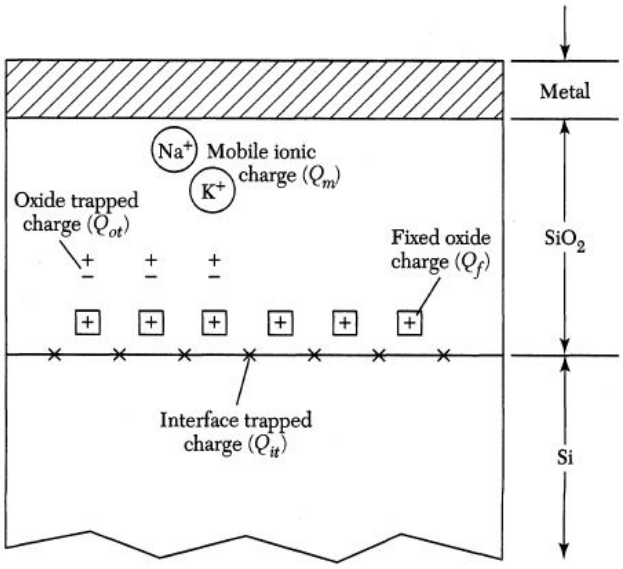
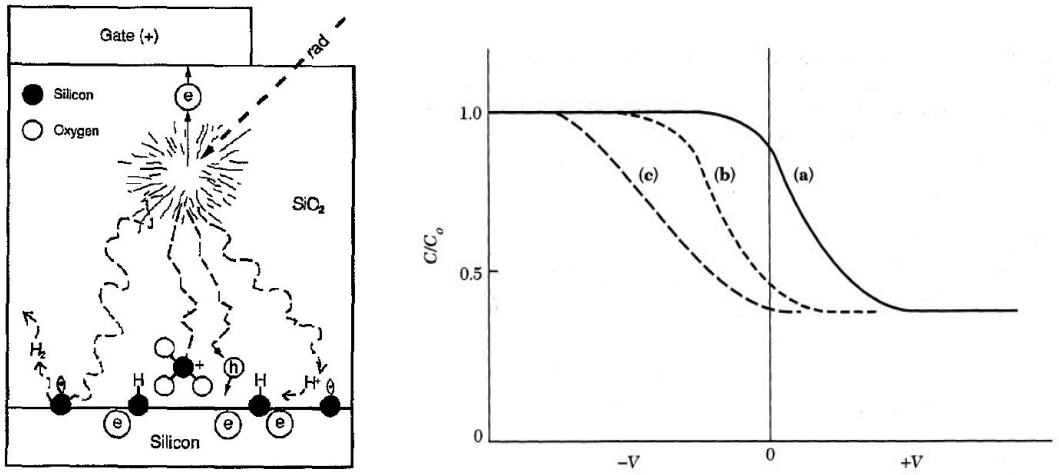


Figure 2.20: Picture of the different types of traps in a SiMOS capacitor: interface trapped charges Q_{it}, fixed traps Q_f, oxide-trapped charges Q_{ot} and mobile ions Q_m [22].

These mean characteristic times can vary from <1 ms to >1 s according to temperature and gate voltage and this is the reason why interface traps show a frequency dependence: at low frequencies $\langle\tau_c\rangle$ and $\langle\tau_e\rangle$ are large enough to play a role in capturing and emitting electrons, whereas at high frequencies the external applied signal is too fast for the activation of these interface states and so their effect is mitigated (see Chapter 6 for the experimental results). Moreover, the fact that these traps are located at different energies within the Si bandgap implies their voltage dependence. This is the main difference between interface traps and the other three types of traps. Their concentration is also dependent on the crystallographic orientation of Si, reaching a minimum value of $10^{10}\ cm^{-2}$ for <100> oriented Si and in the best fabrication conditions, which all include an annealing in hydrogen to passivate the dangling bonds at the interface (see Chapter 6).

The second kind of traps are the *fixed traps* Q_f. These traps are generally positive and dependent on oxide and annealing conditions, because they are constituted by the charge density which remains after interface traps are annealed. For this reason, they are generally located in the oxide within the very first *nm* far from the interface and they are fixed because they do not drift because of external voltage. Also in this case, their minimum value, being again dependent on Si orientation, is in the order of $10^{10}\ cm^{-2}$ for <100> oriented Si. This is the reason why this is the orientation chosen for all the MOS-based devices.

Then, *oxide-trapped charges* are traps created in the oxide because of radiation exposure, like X-rays, high-energy electron bombardment (like in E-beam lithography) and hot carriers injection [37]. In fact, as represented in Fig.2.21a, high-energy radiation can ionize atoms inside the oxide, which can migrate to the interface when positive voltages are applied to the gate. The result is the creation of further interface traps because the migrating atoms can remove the hydrogen which passivate the interface states (this is known as de-passivation).



(a) **(b)**

Figure 2.21: a) Ionization procedure due to high-energy radiation. The ionized molecule is a non-neutral charge which can also migrate to the Si/SiO_2 interface if a positive voltage is applied [38]. b) Ideal C-V curve (a) vs real C-V one without interface traps (b) and with interface traps (c). The last curve is shifted from the ideal value, but also distorted, highlighting the voltage dependence of Q_{it} [22].

In conclusion, the last kind of defects are impurities more than traps: the *mobile ionic charges* Q_m. These ions are typically ionized alkani metal atoms such as sodium or potassium. The main contamination source of these ions is the human body, which regularly produces them. Metal ions can be present at the oxide-semiconductor interface, but as well as the metal-oxide one and their presence can be very detrimental for electronic devices because they easily diffuse through the oxide, especially for devices operating at high temperatures and under large electric fields, shifting the value of the threshold voltage of MOSFETs for example, thus making unwanted conductive or non-conductive connections in the integrated circuit.

Effects of Traps on C-V curves

The presence of traps, being non-neutral charges, shifts the value of the C-V, i.e. capacitance vs voltage curves (see Chapter 6 and Appendix A). The *flatband voltage* V_{FB} is the voltage to apply to the metallic gate of a MOS capacitor to establish a condition in which the bands are flat, i.e. where the net charge at the interface is zero. In an ideal MOS capacitor, V_{FB} is zero, but in a real device there is a difference between the workfunctions of the metal and the semiconductor ϕ_{MS} which makes $V_{FB} \neq 0$. In other words, to compensate the energy difference between the Fermi energy in the metal and in the semiconductor, a net charge $\neq 0$ is created at the interface, responsible for band bending. This means that to recover the flatband condition, a $V_{FB} \neq 0$ has to be applied. And when traps are present, they contribute to this net charge, thus further shifting the value of V_{FB}, as shown in Fig.2.21b and according to [37]:

$$V_{FB} = \phi_{MS} - \frac{Q_f + Q_{ot} + Q_m + Q_{it}(\psi_s = 0)}{C_{ox}}, \tag{2.30}$$

where ψ_s is the surface potential already introduced in Section 2.1.1, C_{ox} is the oxide capacitance and ϕ_{MS} is the workfunction difference between metal and semiconductor. The workfunction is the energy that has to be spent to extract an electron from the highest occupied energy level in the material at 0 K, i.e. the Fermi energy E_F. This formula underlines the fact that only interface traps have a voltage dependence and for this reason in the flatband regime they are evaluated at $\psi_s = 0$ (if the bands are flat, there is no voltage drop on the semiconductor). Moreover, the voltage dependence of Q_{it} is highlighted also in Fig.2.21b: the C-V curve which takes into account them, not only is shifted from the ideal value, but also distorted [22].

2.4 The FD SOI technology

The SET presented in the Thesis is fabricated on fully-depleted silicon on insulator (FD SOI) wafers. SOI are wafers where a *buried oxide*, called BOX, provides an excellent insulation between a thin Si film on the top of the BOX (top Si) and the bulk. In these wafers, devices are fabricated only on the top Si, with the advantages of greatly reducing leakages with the substrate, short channel and body effects (nearly ideal sub-threshold slope) and soft errors due radiation-induced charge generation in the bulk, especially in DRAM cells [39],[40]. Moreover, in high-frequency analog applications, SOI wafers allow to fabricate devices with lower junction areas and so lower parasitic capacitances. And last but not least, if the BOX is thin enough, a back gate can be used to control channel occupancy.
On the other hand, this technology presents also some drawbacks. First of all the

higher cost of the wafers (realized through the *smartcut process* by Soitec [41]), but also source and drain engineering and self-heating effects in the top Si (the BOX makes more difficult heat dissipation).

SOI wafers can be *partially depleted* (PD) or *fully depleted* FD, according to the top Si thickness. In fact, the maximum width of the depleted region $x_{d,max}$ in a MOS device is obtained when the voltage drop on the semiconductor ψ_S is the maximum one, equal to $2\psi_B$ (see Appendix A), and given by [38]:

$$x_{d,max} = \sqrt{\frac{2\varepsilon_{Si}(2\psi_B)}{qN_{top}}}, \tag{2.31}$$

being ε_{Si} the dielectric constant of Si, q the elementary electron charge and N_{top} the doping of the top Si. If the top Si thickness is larger than $x_{d,max}$, it means that it does not matter how large the gate voltage is, the top Si can never be completely depleted, being thicker. In this case, the SOI is known as PD SOI, still affected by short channel effects, as well as by other limitations such as floating body and Kink effect [38]. These limitations are completely overcome in FD SOI, where $x_{d,max}$ is larger than the top Si thickness. Fig.2.22 shows two MOSFETs realized on a PD and FD SOI wafer [33].

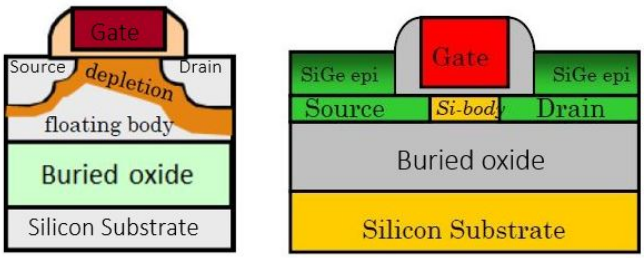


Figure 2.22: MOSFETs realized on a PD (on the left) and FD (on the right) SOI wafer. In the first, a floating body between the depletion region and the BOX is still present, whereas in the second the top Si is so thin to be completely depleted. Nevertheless, raised source and drain contacts are needed to reduce access resistance. Adapted from [33].

2.4.1 Quantum Dots on FD SOI wafers

Being the SOI technology essential for high-performance processors (based on the last generation transistors such as FinFET, TriGate and GAA FET), but also for HF, low-power and high-voltage circuits, optoelectronic devices, MEMS etc.[42], it may be a booster also for quantum technologies. However, currently only very few

implementation of QDs realized on FD SOI wafers exist and all from CEA-Leti in Grenoble [43], [44],[45],[46],[47]. Here, QDs are defined in nanowires, etched in the top silicon, by plunger gates on the their top. Physical etching is a valid alternative to the double metal layer used also in this Thesis to define QDs, because then only one metal gate is needed to accumulate electrons or holes, as shown in Fig.2.23 [46]. However, even though single electron occupancy and manipulation have been

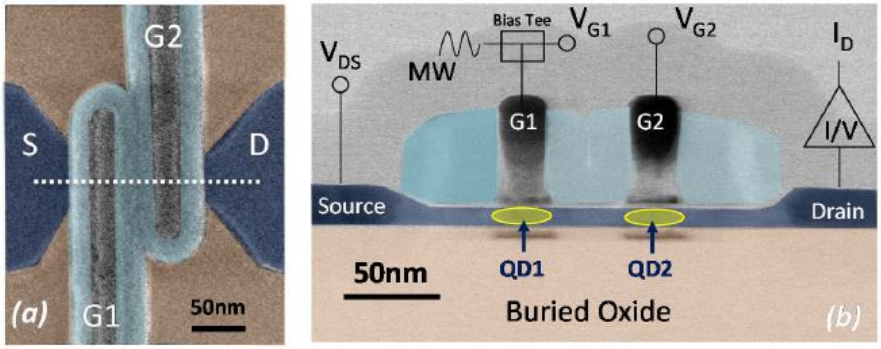


Figure 2.23: Double quantum dot realized in a nanowire physically etched in the top Si of a FD SOI wafer. SEM image of the top view (on the left) and TEM image of the cross section (on the right), showing the position of the two QDs [46].

demonstrated in Ref.[43], [44] and [45], these platforms do not host electron spin qubits. In fact, only hole spin qubits have been realized (see Ref.[46] and [47]), i.e. single-qubit operations like spin flipping have been demonstrated only for holes in the valence band, but still not for electrons. The advantage of realizing spin qubits with holes consists in the possibility of performing single-qubit operation through EDSR, but without the needing of integrated micromagnets. In fact, holes, having a larger effective mass than electrons, exhibit a larger spin-orbit coupling, which in this case can be used as intentional driving mean for coherent hole-spin rotations by applying a time-controlled microwave modulation of the gate voltage [46].
The aim of the *Electron Spin Qubits* project is the realization of high-fidelity electron spin qubits on FD SOI wafers, where spin manipulation is possible through integrated micromagnets combined to voltage signals.

Chapter 3

SOI SET Fabrication Preparation: Process Flow and Cleanroom Tests

Once having understood most of the theoretical aspects necessary for the realization of a SET with SiMOS QDs, the following step is to sketch an ad-hoc process flow for its cleanroom fabrication. This is a step preliminary to device's full layout design, because tests are necessary to understand which are the technologies at disposal for the fabrication of the device, the best materials to use for the different purposes, the critical dimensions achievable with the lithographic tools etc. This is even more important in a Master's thesis project, where the full layout of the device is not optimized from the beginning and, as a consequence, in order not to waste time, preliminary literature study is carried out in parallel with tests in cleanroom, also to get trained on the available tools and get confident with them.

In the contest of this work, as it will be better explained in the dedicated sections, these preliminary tests are:

1. *electron beam lithography (e-beam) resolution tests* to find the minimal dimension of the gates to design achievable with the PMMA resist. These tests include also lift-off tests, since the maximum resolution in the end is given by the pitch of the metal patterned on the wafer and not by the developed resist;

2. on *HSQ* (*hydrogen silsesquioxane*), a negative ebeam resist, used as bond pad protection from wire bonding, to find the best current and dose for its exposure, as well as the correct developing time;

3. *alignment tests*, both for photo and electron beam lithography, to understand

the maximum misalignment in fabricating more layers. The result of this test forced to switch from the more complicated design with three metal layers (plunger, lead and barrier gates) to the simpler with just two, where misalignments of few decades of nm were not deleterious, since at this stage only a proof-of-concept of the SET was necessary and not its integration in the QDs array;

4. *etching tests* to study the etch rate of SiO_2 in HF (*hydrofluoric acid*) and BHF (*buffered* HF), whose results influenced some fabrication steps of the process flow and also the TCAD simulations for the implantation.

Also the study on the gate oxide quality and on the ohmic contacts represent a sort of preliminary tests for the realization of the device. However, they are not entirely cleanroom tests, being the results obtained by outside-cleanroom electrical characterization measurements. Moreover, they represent case studies more than simple fabrication tests and so they are discussed in details in the chapter dedicated to the characterization (see Chapter 6).

3.1 Process Flow

In this section, only the fundamental steps are described in order to understand the choices for the fabrication tests, the TCAD simulations and the layout. Then, the complete fabrication recipe, with all the details, will be illustrated later in the Thesis, when all these preliminary results will be clear.
Before going on, it has to be underlined that the process flow presented in this section is entirely fruit of the work of Fabio Bersano, the PhD student responsible for this project and the in-loco supervisor of this Thesis. Its doctoral program started few months before the beginning of this work, so when this happened the process flow was ready. And even though some details have been modified along the way, all the credits for the presented process flow go to him.

3.1.1 Preliminary Remarks

The initial idea was to fabricate the SET on a FD SOI wafer with a top silicon thickness of about 15-20 nm. This thickness results in the best trade-off between having an UTB easy to fully deplete (i.e. by applying very low voltages) and the realization of good ohmic contacts for source and drain. In fact, one of the problems when dealing with UTB SOI wafers is the realization of ohmic contacts, because the top silicon is so thin that a huge resistance is experienced by the current. For this reason, usually in the last fabrication nodes taking advantage of this SOI technology, the silicon channel is less than 10 nm, but the source and the

drain regions are *raised* to about 20 nm. Usually, these raised source and drain are realized through an epitaxial growth of SiGe, as it is in the SOI qubits published by CEA-Leti.
So, a top silicon thickness of 15-20 nm should have been able to satisfy both the specifications, without the needing of raised source and drain for the contacts.

However, the SOI wafers at our disposal from Soitec had an initial top silicon thickness equal to 70 nm and so a thinning down operation was needed. These operation was realized through an oxidation plus etching procedure. That is because the oxide growth in furnaces is usually very precise, as it is the conversion factor which allows to compute how much silicon is consumed during the oxidation. This factor is equal to ratio between the molecular densities of silicon and silicon oxide [48]:

$$\frac{N_{SiO_2}}{N_{Si}} = \frac{2.305 \times 10^{22} cm^{-3}}{4.992 \times 10^{22} cm^{-3}} = 0.46. \tag{3.1}$$

As a consequence, to thin down the top silicon to the desired value of 20 nm, an oxidation of 100 nm of SiO_2 was asked (oxidation is a service performed by the staff), after having measured the silicon consumed during the RCA cleaning procedure, a procedure mandatory before high-temperature processes. As it will be better explained later, RCA cleaning is a cleaning procedure based on an brief silicon oxidation in a bath and the strip of the created oxide to remove impurities at the surface. So, during this operation, some silicon is consumed because, as just explained, oxide grows at its expense. Ellipsometer measurements revealed a consumed Si thickness of about 5 nm during this operation and so, the asking an oxidation of 100 nm of oxide would have resulted in a final Si thickness equal to:

$$t_{topSi} = (70 - 5 - 0.46 \cdot 100) nm = 19 nm.$$

However, something went wrong with the oxidation and a SiO_2 of 133 nm was grown. The result was a remaining top silicon of just about 7 nm.

This is the reason why two batches of SOI have been processed:

1. one SOI wafer from Soitec with a top silicon thickness of just 7 nm and a BOX of 20 nm;

2. one SOI wafer provided by EPFL's cleanrooms, with a top silicon thickness of 220 nm, thinned down (this time correctly) to a final thickness of 27 nm (after also the second oxidation for the gate oxide), but with a very thick BOX of 2 μm.

Both the two batches consist in p-doped SOI wafers, with a boron concentration from $9.0 \times 10^{14} cm^{-3}$ to $1.5 \times 10^{15} cm^{-3}$ (equivalent to a resistivity from 9 to 15

$\Omega \cdot cm$) for the thinner Soitec wafers and from $1.2 \times 10^{15} cm^{-3}$ to $1.6 \times 10^{15} cm^{-3}$ (equivalent to a resistivity from 8.5 to 11.5 $\Omega \cdot cm$) for the thicker EPFL ones. The advantages of the first batch are the higher quality of the wafer and the possibility of using also the back gate because of the ultra-thin BOX. However, wafers with only 7 nm-thick Si are very difficult to process, because no further Si can be consumed. This forced us to perform a lot of etching tests to thin down the silicon oxide from 133 nm to just 5 nm to use it as gate oxide. Moreover, contacting 7 nm of Si can be very challenging for the reasons already explained (raised source and drain may be mandatory).
On the other hand, the second batch has the advantage that the top Si thickness allows to process it safely and a 5 nm gate oxide can be grown very precisely after having etched away all the silicon oxide resulting from the thinning down procedure. Moreover, there are less problems in the realization of ohmic contacts and also the TCAD simulations for the implantation are easier and show better results. However, the quality of the wafer is, of course, not the same of the ones coming from Soitec and there is not any possibility of using a back gate because of the BOX thickness.

3.1.2 Process Flow Illustration

The main steps for the SET fabrication are now presented, in order to properly introduce the working principle of the tools used during the Thesis and explain the necessity of the fabrication tests illustrated in following section. All the fabrication process is carried out at EPFL *Center of MicroNanotechnologies* (CMi), a complex of cleanrooms and processing equipment.

1) Thermal Oxidation for Top Silicon Thinning Down

SOI wafers have fixed specifications concerning top Si and BOX thickness. So, in order to reach the desired values of top Si thickness, usually a thinning down procedure is needed, procedure which is carried out through an oxidation and the subsequent etching of the grown oxide.

Thermal oxidation is the process through which SiO_2, the natural oxide of silicon, is grown in high-temperature furnaces (800-1200°C). SiO_2 can be deposited through different techniques such as *chemical vapour deposition* (CVD), *physical vapour deposition* (PVD) and *atomic layer deposition* (ALD), but in all these cases it is deposited and not grown. As a consequence, the quality of the deposited oxide is lower, but for its applications as masking layer, passivating layer, sacrificial layer or thermal insulation, it represent the best choice as it can be deposited at lower temperatures. However, for electrical insulation, high quality is needed and this

can be obtained by a thermal oxidation of silicon. Two main approaches exist in this case: wet oxidation and dry oxidation. The first one consists in oxidizing the silicon with H_2O, which is inserted into the chamber under the form of water vapor, whereas in the second one the oxidation takes place because of the presence of O_2. This second oxidation allows to grow an oxide of higher quality with respect to the wet one, but it is slower and so usually the first one is used to grow field oxides (general insulation between the different devices), whereas the second is used for gate oxides, being thinner.
Oxidation is a process which takes place at the Si/SiO_2 interface. This means that, since the new oxide grows at the interface, diffusion of O_2 or H_2O through the already grown oxide plays a fundamental rule in controlling the oxidation rate. As a consequence, two regimes are present during the oxidation process: transport-limited and reaction-rate-limited, as explained by Deal and Groove [49].

Before oxidation, as well as before any high-temperature process, a surface aggressive cleaning, called RCA cleaning, has to be performed on the wafer. RCA cleaning consists in three steps [50]:

1. organic residues removing, through a bath in ammonia and peroxide which oxidize the first nm of silicon. In this way all the organic impurities are transferred in the oxide;

2. thin oxide etching in HF to remove the previously created oxide and also the native oxide;

3. metallic species removal with a hydrochloric acid and peroxide-based bath.

RCA cleaning consumes some silicon, so to properly thin down the top Si of SOI wafers, this effect has to be considered and characterized. A good strategy is to measure the top Si thickness before and after this procedure, so that when the wafer starts the oxidation, the correct thickness is well known.

In CMi's cleanrooms, oxidation is a service performed by the staff, so only the thickness of the oxide to be grown and the technique to employ have to be indicated. For both the two batches of SOI (with 70 and 220 nm of top Si initial thickness), a dry oxidation with DCE (*dichloroethene*, whose benefits will be explained in a dedicated section) was asked, resulting in respectively 133 nm and 390 nm-thick gate oxide. Fig.3.1 shows the schematic of the first batch of SOI wafers from Soitec as arrived (Fig.3.1a) and after the oxidation procedure, i.e. RCA cleaning included (Fig.3.1b). Fig.3.2 shows the same process, but on the second batch of SOI, the ones from CMi. The final top Si thickness is of just 7 nm for the first batch and 39 nm for the second. In the first case, too much silicon was consumed, as already explained several times, and as a consequence, was not possible to completely

Figure 3.1: Top Si thinning down. SOI wafer from Soitec before (a) and after the oxidation (b).

Figure 3.2: Top Si thinning down. SOI wafer from CMi before (a) and after the oxidation (b).

remove the silicon oxide in order to grow, with a successive oxidation, the thin gate oxide. To solve this problem, a series of controlled etching test has been done in order to thin down the silicon oxide from 133 to 5 nm and use it as gate oxide. On the other hand, in the second batch the silicon thickness was the desired one and this allowed to completely remove the SiO_2 and grow a second gate oxide with another cycle of RCA cleaning and oxidation, with a final result of 27 nm of top Si and 5 nm of gate oxide.

2) Silicon Oxide Wet Etching

In order to remove the oxide grown, the simplest choice is a wet etching in fluorine-based chemistry. Etching is the process through which unwanted areas or films are removed. It can be a wet etching, where the material to etch is dissolved in a wet chemical solution (usually an acid bath) or dry, where it reacts with gases (usually a plasma) to form volatile products [51]. Since SiO_2 is an amorphous material, its etching is completely isotropic, which means that there is not a preferred crystallographic orientation etched more than others, like it is in anisotropic etching of silicon for example. As a consequence, the most common technique employed is a wet etching in HF or BHF, also called BOE (*buffered oxide etch*). It consists in a 50% HF and 40% NH_4F solution to obtain a more homogeneous etching because the buffer stabilizes the pH of the bath. HF and BHF etch rates depend on the HF concentration in the solution (from 1 to 50%). For BHF, the HF concentration

in CMi's wet benches is 50%, so the etch rate is very high (>86 nm/min for dry oxide). As a consequence, BHF was used to etch all the silicon oxide from the second batch of SOI wafers (the ones with 39 nm of top Si) and to thin down the first batch from 133 nm to about 20 nm in two steps. Then, the last etching cycle from 20 to 5 nm of gate oxide, always on this firts batch of 7nm-top Si wafers, was performed in HF 1%, which has an etch rate of only about 5nm/min. The results are shown in Fig.3.3 for both the two batches.

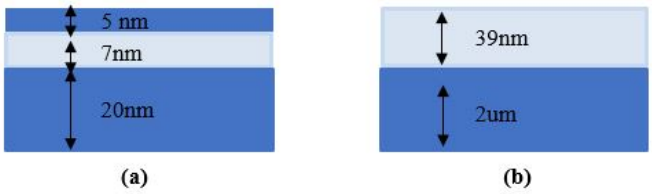


Figure 3.3: SOI wafers after SiO_2 wet etching in HF/BHF. a) SOI wafer from Soitec (BHF + HF 1% etching to get the 5 nm-thick gate oxide. b) SOI wafer from CMi (only BHF to completely remove the oxide).

3) Gate Oxide Thermal Oxidation

This step concerned only the batch with 39 nm of top Si, i.e. the only one where there are no problems in consuming silicon with the oxidation step. After having removed all the SiO_2 in BHF, the wafer was first RCA cleaned. This time, the as mentioned step consumed about 10 nm of Si, leaving just about 29 nm. The wafer was then put into the furnace for the oxidation of 5 nm-thick gate oxide (dry oxidation with DCE at 850°C). After the oxidation, again ellipsometry measurements revealed a remaining top Si thickness of 27 nm, as expected, since the oxidation of 5 nm of SiO_2 consumed: $Si_{consumed} = (0.46 \cdot 5)nm = 2.3nm$. Fig.3.4 shows the result of this last oxidation step.

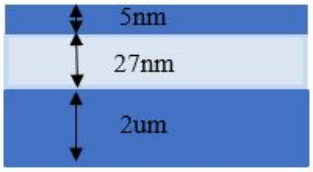


Figure 3.4: Gate oxide thermal oxidation result on the CMi's SOI wafer. The gate oxide is 5 nm-thin and the top Si final thickness is of 27 nm.

4) EBL, Metal Deposition and Lift-off for Markers

At this point of the process flow, the two SOI batches have finally the same gate oxide thickness. The next step is the realization of metallic markers for two purposes: the first set consists in big crosses (1.5 mm × 1.5 mm) which define the borders of the different chips in which the wafer will be diced (dicing markers), whereas the second one in squares of the dimension of 20 μm × 20 μm which are used by e-beam lithography software as references for the alignment of the subsequent patterned layers (EBL markers). These markers are patterned with a heavy metal (usually tungsten, but also platinum, like in this case) to be more easily recognised by the ebeam tool (heavier metals give a better contrast). A more detailed description of this markers will be given in Chapter 4, the one dedicated to the design of the device and the layout of the different layers.

To pattern desired shapes at the micro and nano scale, lithography is the approach to be used. Lithography, as the name suggests, is the process which allows to transfer desired patterns to a certain layer. In *photolithography* (PL), an organic material, called *photoresist* (PR), is selectively exposed by light through a mask which is transparent in the regions to be exposed and opaque everywhere else. In this way, the pattern is transferred from the mask into the PR, which can be then more soluble in a chemical solution called *developer* if it is a *positive* PR, or less soluble if it is a *negative* one. In the first case, the pattern defined in the mask will be transferred to the PR after the development, whereas in the second its complementary (that is why in this case the PR is called negative). Fig.3.5 [52]

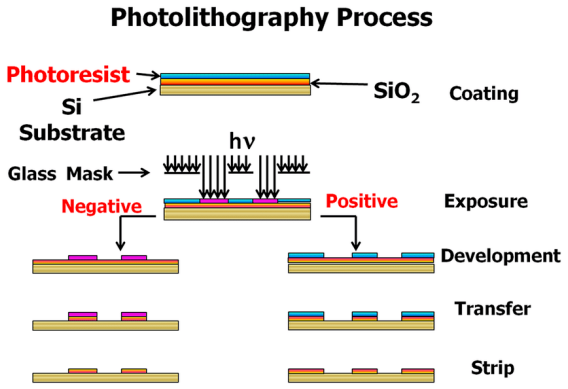


Figure 3.5: Standard PL process: 1) resist spin coating 2) exposure 3) development 4) transfer of the pattern onto the material (through etching in this case) 5) resist strip [52].

shows the complete process just described, which includes also the PR *spin coating*, transfer of the desired pattern (through etching, if the material to pattern is already

present before PL, so that the PR acts as a mask to etch the material only in the desired regions, or *lift-off*, if the material is deposited onto the developed PR, so that only the material deposited on regions in which the PR is not present will survive to process), development in a chemical solution to remove only the exposed resist (if positive) or the not-exposed one (if negative) and resist strip, the action which removes the organic PR from the sample. However, in PL the use of a mask is not mandatory. Some *direct laser writing* technologies are available, where a laser exposes the resist only in the regions of the pattern (or the complementary ones in case of negative PRs). For this kind of PL, the preliminary fabrication of a chromium mask is not necessary, but the software of the tool is able to convert the *.gds* file of the layout directly into the *.job* file which tells the laser exactly where to write according to the designed layout. An example of this technology is the MLA150, from Heidelberg Instruments [53], available in CMi's cleanrooms and used in this work for the etching of the thick Al_2O_3 in the dot region, for the final re-opening of the metal pads buried under the oxide layers at the end of the process, as well as for some MOS capacitors used for the study on the gate oxide quality. On the other side, the drawback, of course, is that this kind of technique is slower with respect to standard PL, where the whole pattern is exposed simultaneously and not serially pixel by pixel.

This same drawback is shared with *e-beam lithography* (EBL), where a focused beam of electrons is used to expose the resist instead of light. Since electrons have a wavelength (the de Broglie wavelength) of about 0.2-0.5 Å [51], a nm-scale resolution can be achieved. In e-beam lithography, electrons are usually generated by a thermal field emission gun and accelerated to the sample by voltages that can go up to 100 keV. The electron beam is deflected by magnetic coils, but since the maximum deflection is usually very small (about 100 μm [54]), it is the stage which moves to draw the entire pattern once the maximum deflection is reached. The schematic of the e-beam column is depicted in Fig.3.6a, whereas Fig.3.6b shows the Raith EBPG5000 ebeam tool in the dedicated cleanroom in CMi. Raith

(a) **(b)**

Figure 3.6: Electron beam lithography column (a) and CMi's Raith EBPG5000 ebeam tool (b).

EBPG5000 allows to reach resolutions <20 nm, by properly choosing the resist, the current of the beam (which defines the dimensions of the beam spot) and the dose, i.e. the number of charges per unit of area, given by: $dose = \frac{current}{f \cdot area}$, being f the frequency with which the beam is stepped from one pixel to another when writing a pattern (up to 50 MHz in this case). Further considerations on these parameters will be discussed in Section 3.2.1, the one dedicated to the resolutions tests.

It was decided to use EBL also for the patterning of these markers, even though they are not so small, because markers for EBL need to be very sharp and precise with a small edge roughness to reduce misalignment errors. In Fig.3.7 is illustrated the process of exposure and development of the double layer resist stack used for this step. It consists in a first layer of MMA (*methyl methacrylate*) and a second layer of PMMA (*poly methyl methacrylate*), the most common e-beam resist. PMMA is a polymer constituted by monomers of MMA, so it is more robust than just MMA and, as a consequence, exposed regions in MMA develop more than the same in the PMMA layer in the developer solution. The result is a "T" structure, useful to facilitate the lift-off process after the deposition of the metal layer. A good "rule of thumb" for a good lift-off process which will be better explained in a little while, is to make the MMA layer with, at least, a double thickness with respect to the thickness of the material to deposit. After the lithography, the following step is the

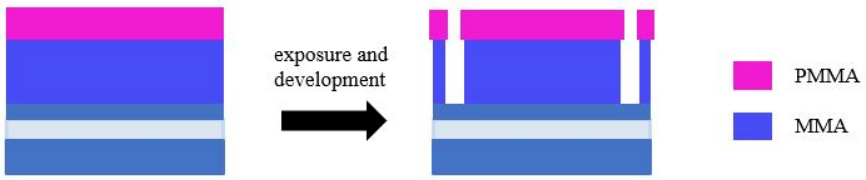


Figure 3.7: E-beam lithography exposure and development of the double layer MMA/PMMA resists. The "T", highlighted by the undercut of the first resist layer, helps the lift-off process after metal deposition.

metal deposition for the markers. Different deposition techniques are available for metals, but the most common ones are *sputtering* and *evaporation*, which both fall in the PVD techniques for thin films. Since all the patterned metal layers of the SET of this work are obtained with lift-off, the second option was chosen because the long distance (1 m) between the crucible, where the target to deposit is host, and the sample, combined with the "T" structure allows a metal deposition just in the holes, greatly limiting the one on the sidewalls of the developed resist [51]. In fact, the lift-off process consists in removing all the resist in a wet chemical solution, so that also the metal deposited onto the resist will be removed, whereas the one deposited in the developed regions will survive. For this reason, deposition

also on the sidewalls can give origin to a continuous thin metal film that is stripped all together when the resist is removed. In Fig.3.8 [55] is represented the difference between etching and lift-off. Coming back to the metal deposition technique, the

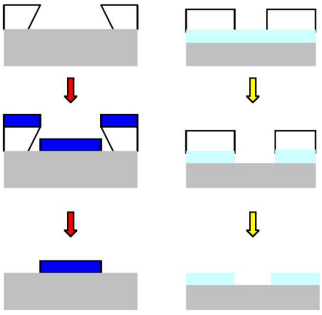


Figure 3.8: Lift-off (on the left side) vs etching (on the right side). The fundamental difference is that in the etching process the material to pattern is deposited before the lithographic step and the patterned resist plays the role of mask to protect the regions covered from resist from etching. In lift-off the material is deposited after the lithographic step, so that when the resist is stripped away, only the metal deposited in the developed regions will survive [55].

evaporator used for all the metal layers is the LAB 600H from Leybold Optics, an e-beam evaporator. In e-beam evaporators, as shown in Fig.3.9 [56], an electron beam is accelerated and deviated toward the crucible. The heat generated make the material melt and evaporated toward the sample. The figure highlights also the

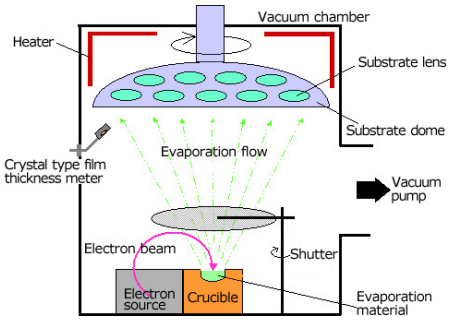


Figure 3.9: E-beam evaporator schematics [56].

presence of a vacuum pump because to deposition occurs in high vacuum (about 1.5×10^{-6}), but multi-layer deposition are possible without breaking the vacuum because the machine can host more than one crucibles. This is useful also in the context of the SET realization, since the Pt used for the ohmic contacts and the Pd for the gates are all deposited after the deposition of a thin Ti layer, used to promote the adhesion to the substrate. Deposition end thickness is monitored

through the change in resonance frequency of a planar quartz resonator inside the chamber.

After the explanation of how metal patterns are defined in the SET process flow, the steps related to the markers realization which follow the resist development (Fig.3.7), are summarized in Fig.3.10. The metal deposited is Ti/Pt, because Pt is

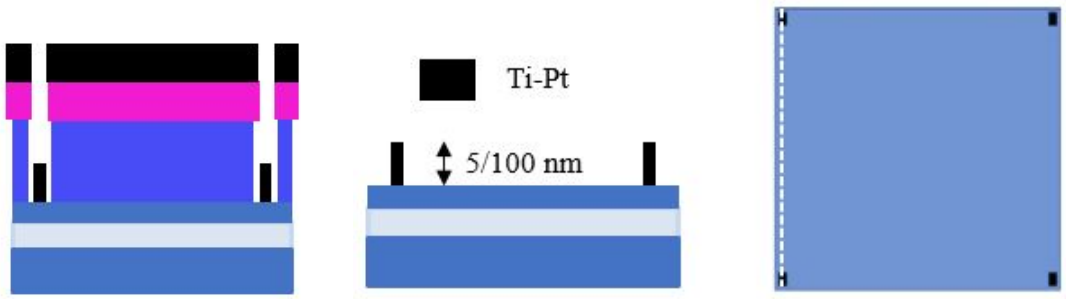


Figure 3.10: Ti/Pt markers patterning. a) Metal deposition (e-beam evaporation) and b) lift-off (cross section and top view).

a heavy metal (large atomic number) which gives, being also relatively thick (5/100 nm), a good contrast. In this way, it is easily recognized by the EBL tool when it tries to locate the alignment markers.

5) EBL for Ohmic Regions (n+ Source and Drain)

A second e-beam lithography step is required to define, still at the wafer level, the ohmic regions, heavily n-doped regions where the source and drain metal-semiconductor contacts are realized. These contacts, in order to be ohmic (i.e. linear relationship between the current and the voltage drop, with a very low contact resistance) need a heavily doped silicon ($> 10^{19} - 10^{20} cm^{-3}$) in the regions where the metal contact is deposited. The doping procedure can be done either by *diffusion* or *implantation*. In the first case, dopants present in a top deposited layer diffuse into silicon because of the high temperature. However, this approach was excluded because in SOI wafers, the top Si can be so thin that lateral diffusion of dopants cannot be neglected. If this happens, dopants will contaminate also the regions where they are not supposed to be and, being the doped regions relatively close to the QD, the risk of doping the whole conductive channel under the top gate is very high. To overcome this problem, doping by implantation is typically chosen. However, the minimum implantation energy in conventional system is 5

keV, which is too high to implant UTB SOI wafers without damaging the BOX (see Chapter 4.2). As a consequence, *plasma-immersion ion implantation*, or simply *plasma doping*, is chosen. In this system, used for very high doses ($> 10^{16}$ cm^2) and low energies (< 5 keV), the ions in the created plasma are accelerated to the wafer surface by short (in the order of μs) negative potential pulses applied to the wafer platen at energies up to the maximum desired value. However, this doping technique presents some drawbacks such as an higher cost, potential damage to the BOX because of the high voltages applied to the wafer and sputtering of silicon instead of doping for high doses [57].
In order to implant only in the ohmic regions, a mask has to be patterned to protect all the other regions of the wafer. This mask consists in a MMA/PMMA double layer of respectively 200 and 100 nm, patterned by EBL. Fig.3.11 shows the

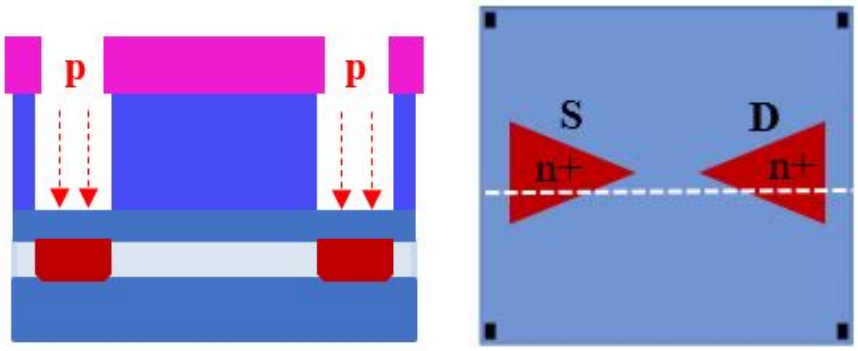


Figure 3.11: Ohmic regions. a) Cross section and b) top view (after resist strip).

simplified cross section (Fig.3.11a) and top view (Fig.3.11b) of the SOI implanted wafer. In the case of the second one, the resist used as mask has already been removed. Resist strip after implantation requires a combination of wet chemicals and plasma oxygen treatment, because usually the resist is burnt because of the ion bombardment. Wet chemicals are the classic baths used also for resist strip during lift-off, whereas the second one is a dry etching technique based on plasma constituted of oxygen radicals which react with organic materials, removing any residue. Oxygen plasma treatment is used also for surface preparation before spin coating to improve adhesion (in the Tepla GiGAbatch tool), but also, at a lower power and for just 30 seconds, to remove resist residues after every development step of the process flow (Oxford PRS900 tool).

6) Dicing

The implantation is the last step led at wafer level, since the the ion implantator tool is not available in CMi and so a shipping to an external fab is needed. After the resist strip step described in the previous step, the SOI wafers are sent to dicing to get 1 cm × 1 cm chips. Dicing is a service performed by CMi staff, who follows the big crosses patterned before to dice the wafers into chips. Before this operation, a surface protection is needed and this can be easily obtained by spinning a thick photoresist onto all the top side of the wafer. The tool used for dicing is the Disco DAD321, which basically consists in a diamond blade (see Fig.3.12 [58]) through which very sharp edges can be obtained thanks to the presence of diamond particles embedded in a bond matrix.

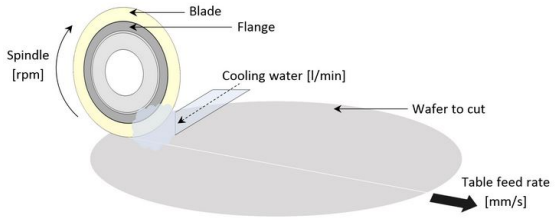


Figure 3.12: Dicing tool schematic [58].

Rapid Thermal Annealing for Dopants Activation

One of the drawbacks of doping by implantation is that to recover the crystalline structure of silicon and to activate dopants (i.e. to collocate them in the *substitutional sites*), an high-temperature annealing is necessary. Long annealings can be done in furnaces like the ones used for oxidation or CVD, but in the most recent technologies *rapid thermal annealing* (RTA) is preferred, because during the annealing process dopants diffuse and long annealing times can result in large diffusion lengths. In fact, *rapid thermal processing* tools, like CMi's JetFirst200, allow to reach temperatures of 1000°C in few seconds thanks to a lamps based system which transfers heat by electromagnetic radiation. With this mechanism, thermal gradients of about 150°C/s are possible, so that RTA usually last from 5 to 30 s if temperatures between 900 and 1000°C are employed [51]. Fig.3.15 [59] eillustrates the schematic of the RTP machine used in CMi, which includes the two temperature monitoring tools (the thermocouple and the pyrometer, the latter used for temperatures higher than 800°C) and a vacuum system to pump the vacuum and then purge into the chamber different gases according to the process (N_2 and forming gases, i.e. 95% N_2 and 5% H_2, for RTA, but also O_2 for RT oxidation and NH_3 for RT nitridation).

This annealing step has been intensively studied during the Thesis to understand

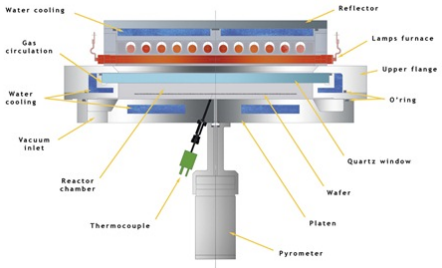


Figure 3.13: RTP JetFirst200 tool schematic [59].

the effect of the highest-temperature step of the process on the quality of the gate oxide (see Chapter 6.1). Under the light of these results, the chosen annealing temperature is 900°C for 20 s for the Soitec SOI chips and for just 5 s for the CMi thicker ones. Since this step is probably the most crucial of the whole process flow, it is performed at chip level and so after the dicing procedure.

7) EBL, Deposition and Lift-off for Ohmic Contacts and EBL Markers

The same metallization procedure used to pattern the first markers is used for the Ti-Pt source and drain ohmic contacts, as well as for a new set of Ti-Pt markers for the align of the subsequent layers. In fact, after the 1000°C RTA, the markers patterned at the first stage can lose their sharp edges). After PMMA exposure, development and post-exposure baking, the chip is dipped into HF 1% to remove the gate oxide in the metal-semiconductor contact regions (a diluted concentration of HF and a post-exposure baking are needed in order to use the PMMA resist as an etching mask). After this etching step, the sample is quickly transferred into the e-beam evaporator to start the deposition the regrowth of a thin native oxide. Metal thicknesses of 5/55 nm are chosen respectively for the Ti adhesion layer and the Pt layer. In Fig.3.14a are illustrated the EBL exposure and development, whereas in Fig.3.14b the cross section (on the left) and top view (on the right) of the deposited contacts after lift-off.

8) EBL for Bond Pad Protections

In order to protect the thin gate oxide from breaking during the wire bonding, necessary to mount the sample into the cryogenic setup, 120 nm-thick HSQ bond pad protections are patterned under the metallic pads of the gates. Only the metallic gates have the oxide below, whereas in the ohmic contacts the oxide is removed just before their deposition. HSQ is a negative EBL resist which, once exposed, remains on the sample and successive exposures further strengthen its

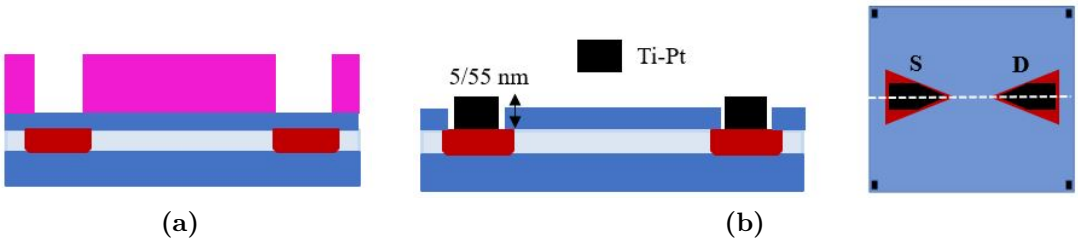


Figure 3.14: Ohmic Ti-Pt contacts. a) PMMA after exposure and development. b) Cross section (on the left) and top view (on the right) of the deposited contacts after lift-off.

bounds. Only HF removes HSQ, which is basically liquid SiO_2, but HF is not used anymore in the process flow.

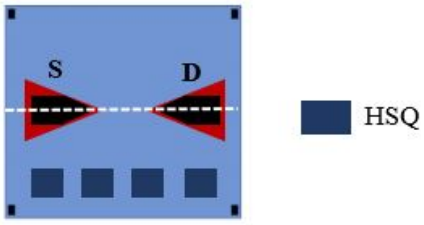


Figure 3.15: Top view of the patterned HSQ bond pad protections after exposure and development.

9) ALD for Al_2O_3 Blanket Oxide and QD Window Wet Etching

A gate oxide of 5 nm is not sufficient to ensure no leakages if also long routings from the metal pads to the dot region are taken into account. To solve this problem, 10 nm of Al_2O_3 are deposited on the whole chip through *atomic layer deposition* (ALD). In this technique, sequential precursor gas pulses are released into a vacuum chamber to deposit one layer at a time on the substrate [60]. Two gases are introduced in each cycle, i.e. a cycle is made up of two pulses: the first one is the *precursor* of the material to deposit, which produces a monolayer on the wafer surface, whereas the second one is the *co-reactant*, which reacts with the first precursor to produce a monolayer of film on the wafer surface. Since each pair of gas pulses (one cycle) produces exactly one monolayer of film, the thickness of the deposited film can be precisely controlled by the number of cycles. The working principle is the following:

1. the first precursor is pulsed into the reactor and reacts at the surface of the sample by chemisorption to form a monolayer;

2. the reaction chamber is purged with inert gas (N_2) and the excess of the precursor is removed.

3. the second precursor (the co-reactant) is pulsed into the reactor and reacts with the first precursor already present at the surface of the sample. The reaction generates the desired monolayer and a byproduct;

4. the byproduct and the excess of precursor are then purged from the reactor with inert gas (N_2).

For Al_2O_3 the gas precursor is *Tri-Methyl Aluminum* (TMA), whereas the co-reactant is H_2O.
However, in this way the oxide in the dot region becomes thicker and, as already anticipated, some works ([21], [20]) show that a reduction of the charge noise is demonstrated for thinner oxides. For this reason, the Al_2O3 blanked oxide is etched away in the dot region (10 μm × 10 μm), leaving only the thin SiO_2 gate oxide. This process is possible through a lithographic step to define the dot region (this time, standard lithography is used, i.e. the direct laser writer MLA150 tool) and the use of H_3PO_4 (*orthophosphoric acid*), which selectively etches Al (and so Al_2O_3) and not SiO_2 if heated-up to 60°C, as sketched in Fig.3.16b.

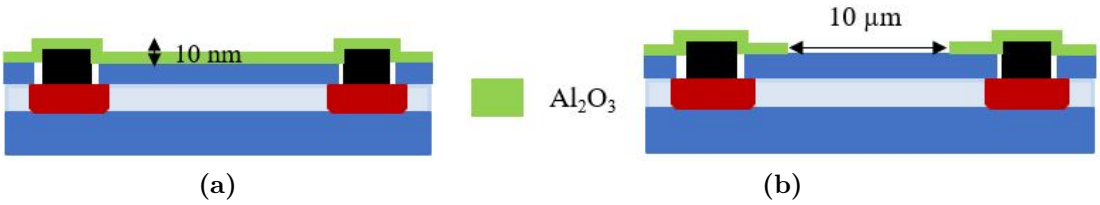
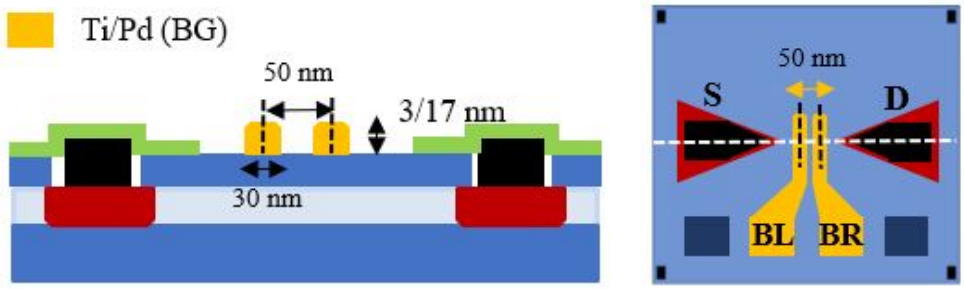


Figure 3.16: Al_2O_3 ALD for blanket oxide, as deposited (a) and after its etching in the dot region.

10) Gates patterning and interlayer oxide

At this point of the process flow, everything is ready for the patterning of the two gate layers which define the dot: the top gate to create the conductive channel and the two barrier gates to locally deplete the channel below then, thus creating the QD between them. The first gates to be patterned are the barrier gates, always through the already illustrated sequence of EBL, metal deposition and lift-off. The metal chosen for this purpose is Pd because of its better uniformity eith respect to

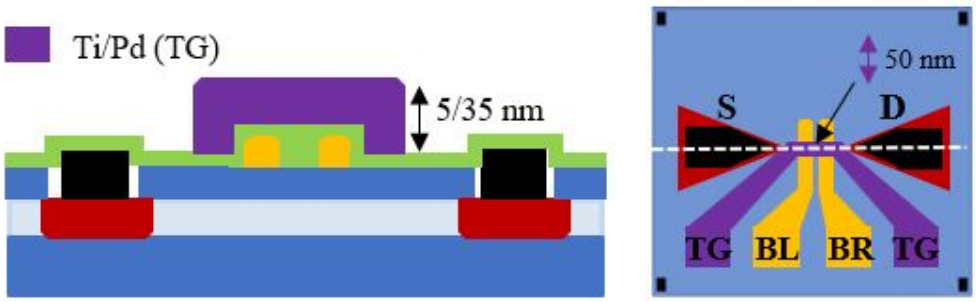
the traditionally used Al [61], deposited on few nm of the usual Ti adhesion layer. For the barrier gates, being the first metal layer above the dot, a thickness of just 3/17 nm of Ti/Pd is chosen. The schematic of the result of this fabrication are shown in Fig.3.17. In order to improve Pd uniformity [62] and reduce interface



(a) **(b)**

Figure 3.17: ELB, Ti/Pd deposition and lift-off result for the first layer of barrier gates. a) Cross section highlighting the width of the barrier gates and of the dot. b) Top view of two barrier gates: left (BL) and right (BR).

traps, a RTA at 400°C for 15 minutes is done in forming gases. Then, 5 nm of Al_2O_3 are deposited by ALD to electrically insulate the two metallic layers (interlayer oxide). After that, the top gate is patterned, always following the same fabrication step of the ones used for the other metal layers. This time, the metal thickness is of 5/35 nm, since no other layers are patterned above and so steps due to underneath layers thickness is not a problem (Fig.3.18). Also after this Pd deposition step, the



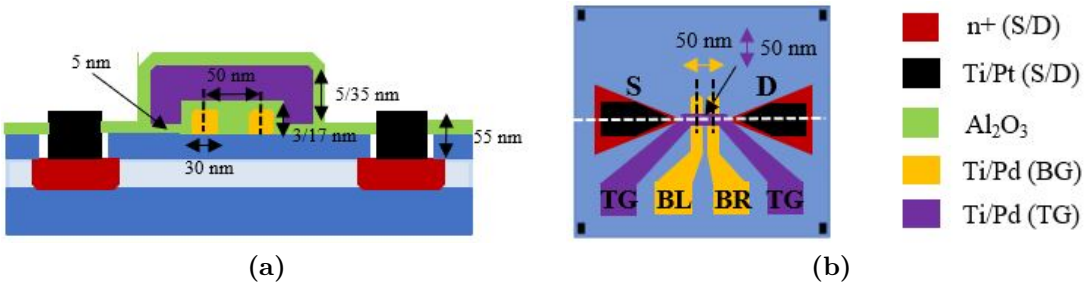
(a) **(b)**

Figure 3.18: ELB, Ti/Pd deposition and lift-off result for the second layer of top gate. a) Cross section. b) Top view.

sample is annealed for 15 minutes in FG at 400°C.

11) Passivation Layer, Contacts Opening and Wire Bonding.

An Al_2O_3 passivation layer of 10 nm to package and protect the device. Contacts are re-opened through the use of MLA150 direct laser writer to expose the resist only in the large 100 μm × 100 μm pads regions, so that an etching step in H_3PO_4 removes only the oxide above them, allowing the electrical contact. Electrical contact which is realized through a wire bonding procedure, in order to safely mount the chip in the little sample holder of a dilution fridge for cryogenic measurements. This would be the correct measurement setup for charge noise characterization (mK temperatures), but preliminary measurements at 4 K, i.e. just in liquid He, are sufficient to demonstrate the single electron regime (*Coulomb blockade*).



(a) **(b)**

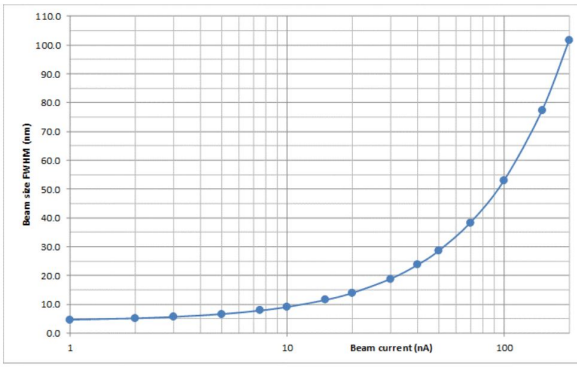
Figure 3.19: Overview of the SET after the last fabrication processes (passivation and contacts opening). a) Cross section with all the dimensions of the different deposited layers. b) Top view of the SET, indicating the dimensions of the dot: 50 nm × 50 nm.

3.2 Fabrication Tests

Having illustrated all the steps of the process flow, as well as a quick overview on the technology used to fabricate the SET, the preliminary tests prior to the full device fabrication are now illustrated. Again, as already anticipated, also the fabrication of the MOS capacitors for the study on the gate oxide and the fabrication of ohmic contacts on UTB SOI chips at disposal of the lab (unusable for the SET being the whole top Si heavily doped) represent studies preliminary to the device fabrication. In fact, the results of the first, allow to understand the best parameters to fulfill for a good oxidation and the influence of the annealing time and temperature for the dopants activation on the oxide quality, whereas the seconds give information about the best choice of the material to realize the contacts and the minimum doping to target in the TCAD simulations for implantation.

3.2.1 EBL and Lift-off Resolution Tests

E-beam lithography is the most used tool for the fabrication of the SET. However, it is also the most complex one to use, also because in EPFL EBL is not a service taken in charge by the staff, but users, after a first quick training, have to solve all the problems occurred along the way by themselves. EBL is a complex tool starting from data preparation: first, the layout is realized in a dedicated software able to produce a *.gds* file (KLayout in the context of this Thesis). Then, the conversion into the *.job* file read by the tool is not immediate like it is for the MLA150 direct laser writer, but the *.gds* file has to pass through a software which converts it into a *.gpf* file. Only this format is understood by CJOB, the software used by the e-beam tool. For the first conversion, two software are available: CATS and BEAMER. Both play the role of fracturing the design into pixels of the desired dimensions (usually about 5 or 10 times smaller than the critical dimension of the design for good results), but BEAMER is used also for advanced functionalities like proximity error corrections and writing order control. Being these functionalities not necessary for the SET, CATS was initially chosen, being the one used also in the company environment. However, BEAMER was then adopted to solve the alignments problems arose from file conversion in CATS.
The dimension of the pixels in which the design is fractured is called resolution and it is a fundamental parameter to pattern features with a very small critical dimension, like it is for the SET. Gates of decades of nm require a pixel resolution of 5 nm (the minimum one) and an appropriate beam current able to generate a beam spot of about 5 nm. In fact, since the beam hits the resist once for each pixel, a good writing strategy consists in choosing a beam with a spot a little bit larger than 5 nm in order to create a smooth pattern. The relationship between the beam current and the corresponding beam size if illustrated in Fig.3.20 [54]. However, the beam



Beam current [nA]	Beam size [nm]
0.1	4.2
0.2	4.2
0.3	4.3
0.4	4.3
0.5	4.4
0.7	4.5
1	4.6
2	5.1
3	5.6
5	6.6
7.5	7.8
10	9.0
15	11.5
20	13.9
30	18.8
40	23.7
50	28.6
70	38.3
100	52.9
150	77.4
200	101.8

Figure 3.20: Beam current vs beam spot: graph and table of the values available with the Raith EBPG5000 tool [54].

current, which determines the size of the beam, and the pixel resolution are not the only two parameters which determine the critical dimension that can be achieved. In fact, the dimensions of the features patterned at the resist level depend also on the dose of the beam, i.e. how many electrons expose the resist per unit of area (see the previous section). Like in standard photolithography, there is a threshold below which the resist is not completely exposed (underexposure), but also higher doses than the threshold can be a problem (overexposure). For EBL, since the beam is gaussian, doses higher than the optimal value result in larger features printed on the resist. In the worst case, when there are two features very closed separated only by a narrow gap, overexposure of both results in the conjunction of them (the gap disappears). And this is exactly the case of the two barrier gates of the SET, separated by just 30 nm in the final design. For this reason, a resolution test is mandatory to find the proper currents and doses to properly pattern the SET.

The software which takes the fractured *.gpf* file generated by CATS and allows to choose the proper beam and dose for each layer to print is CJOB, which combines these information (fractured design, beam current and dose) to generate the *.job* file read by the ebeam tool.

1) MMA/PMMA Bilayer and "Bigger" Beam Currents

The first test with EBL are carried out using the double MMA/PMMA layer. This is because, as already explained, to ease the lift-off procedure an undercut is preferred and this is possible using two resists with different molecular weights: the lighter one as first layer (so that is develops more) and the heavier one on its top. For the very first test, MMA EL6 and PMMA 495K A2 are used, with a respective thickness of 100 and 45 nm. Three different currents of 100, 70 and 15 nA, equivalent to a beam spot of respectively 52.9, 38.3 and 11.5 nm, are used to pattern arrays of two parallel lines with 200, 100 and 50 nm separated between them by the same distance of their width (so, 200, 100 and 50 nm). Doses from 800 to 1200 $\mu C/cm^2$ are used, being the standard dose for PMMA about 1000 $\mu C/cm^2$. The results of this first test show as follows:

1. 200 nm lines are successfully patterned for every current and dose;
2. 100 nm lines are successfully patterned for the smallest current for every dose and for the second current for doses lower or equal than 1000 $\mu C/cm^2$, but they fail for higher doses and for every dose if the biggest current is used;
3. 50 nm lines are always joined together for every current and dose.

Fig.3.21 shows the SEM (*Scanning Electron Microscope*) results of one of the cases: 70 nA current and 1000 $\mu C/cm^2$ dose. This first test shows that up to 200 nm of

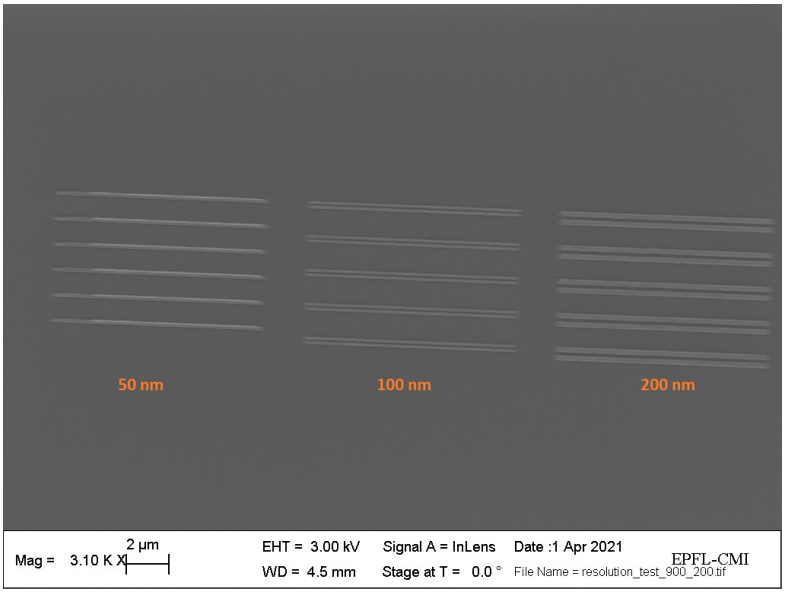
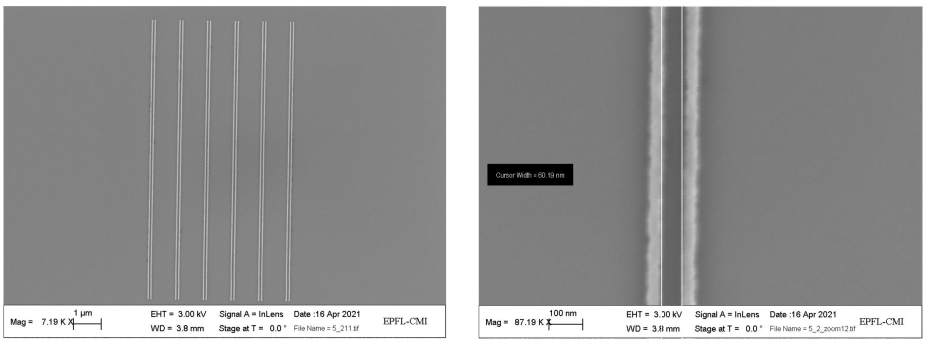


Figure 3.21: SEM picture of the 50, 100 and 200 nm Ti/Pd lines patterned with a beam current of 70 nA and a dose of 1000 $\mu C/cm^2$.

critical dimension, it is very easy to get good features independently on the beam current and the dose. These parameters start to play a role for features of 100 nm, but they fail completely for 50 nm ones. This suggests that smaller currents have to be used for the following tests. The other hint for the following tests consists in the role of the dose, understood considering the 100 nm-features: for higher doses, the overexposure of the resist makes also the gap between the two lines exposed, so that they join together. As a consequence, in the following tests, also doses lower than 1000 $\mu C/cm^2$ have to be chosen.

2) MMA/PMMA Bilayer and "Smaller" Beam Currents

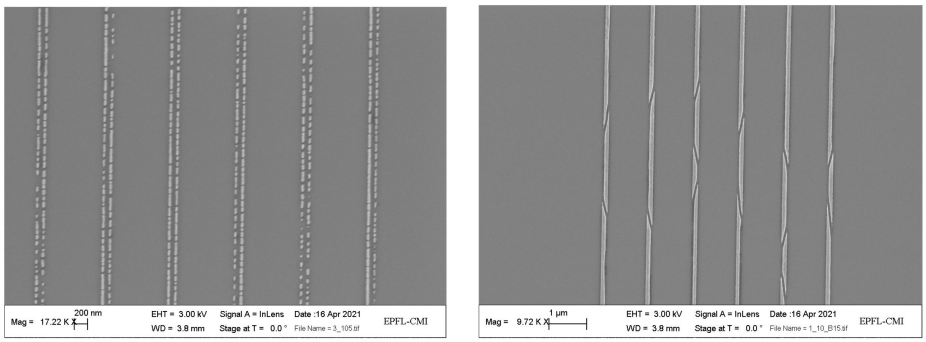
A second test focuses just on 50 nm features, being the SET gates equal or even smaller than this size. This time, four different currents are used: 10, 7.5, 1 and 0.5 nA, giving a beam size respectively of 9, 7.8, 4.6 and 4.4 nm. Finally the dimension of the beam better fits the one of the pixels in which the design is fractured (the resolution in the *.gpf* file). Doses from 500 to 1000 $\mu C/cm^2$ are chosen. The wanted result is reached, but only with the two smallest beams and a dose of exactly 550 $\mu C/cm^2$, as shown in Fig.3.22. Lower doses than the optimal one result in broken and not continuous lines (Fig.3.23a), whereas for higher doses the two lines join together and the gap between them disappears (Fig.3.23b). Moreover, the joined lines are broken in points which, if linked, build a sinusoidal curve. This effect is due to the ultrasounds used to help lift-off, which are acoustical sinusoidal



(a) **(b)**

Figure 3.22: SEM image of the successfully patterned 50 nm gates, for a current of 0.5 nA and a dose of 550 $\mu C/cm^2$. a) Overview of the couples of lines. b) Zoom which underlines the gap between the two lines.

waves. Usually some minutes of sonication really help lift-off, but this time the effect was weird. In fact, it can be noticed from Fig.3.23 that the cracks are not casual, but follow a sinusoidal shape with a wavelength which perfectly matches the frequency of the ultrasounds. The results of this test finally allow to fabricate



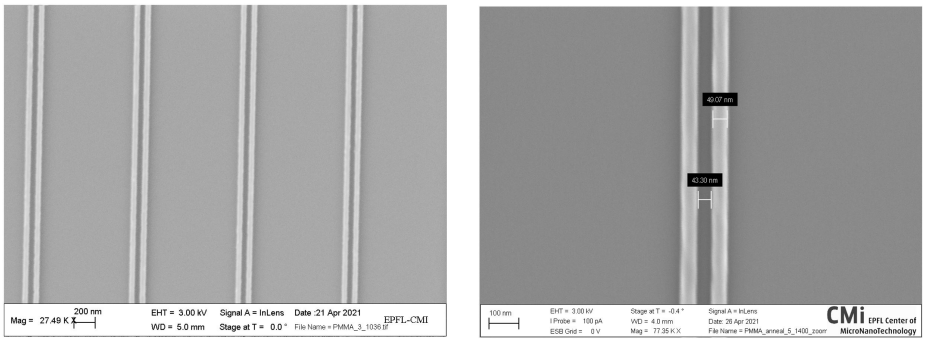
(a) **(b)**

Figure 3.23: SEM picture of the underexposed (a) and overexposed (b) Ti/Pd lines. The dose in the first case is of 500 $\mu C/cm^2$, 900 $\mu C/cm^2$ in the second.

50 nm gates, but the fact that only a dose allows to do it, together with the fact that the measured gap between the two lines is 60 nm and not 50 nm as expected, are sufficient to move toward other resolution tests, in order to improve the yield of the successfully patterned gates and push also toward lower critical dimensions.

3) Heavy PMMA for 50 nm-pitch features

MMA/PMMA bilayer resist usually offers the best results in terms of lift-off, but when the features are too small and specially very closed each other, the undercut in the MMA layer can be so enhanced from both sides that the heavier PMMA on the top can make the structure collapse. For this reason, for this new test only PMMA is used and with a higher molecular weight with respect to the previously used one. This is because for lift-off again the resist should have a thickness which is at least the double of the metal to deposit, so a heavier PMMA allows to obtain a thicker resist after spin coating. Moreover, in general, resists with higher molecular weights allow to achieve higher resolutions [63]. For these reasons, PMMA 950K A2 is used for this test and, in the end, it will be the resist used also for the realization of the SET gates. The thickness of the spun PMMA is between 90 and 100 nm, so again the double with respect to the Ti/Pd deposited. Again, the same four beam currents of the previous test are used: 10, 7.5, 1 and 0.5 nA, with doses from 500 up to 1400 $\mu C/cm^2$ (higher than before since the PMMA has an higher molecular weight). Fig.3.24 highlights that this time the yield is almost perfect:

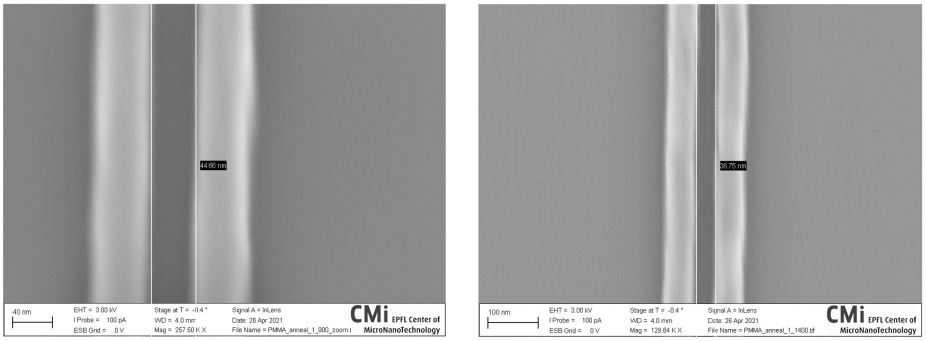


(a) **(b)**

Figure 3.24: SEM image of the 50 nm lines patterned with a current of 1 nA and a dose of 700 $\mu C/cm^2$. Overview (a) and zoom which shows the 50 nm lines separated by about 45 nm of gap (b).

the 50 nm lines are successfully patterned for all the currents and doses. This suggests that the choice of the resist is a key parameter, together with the beam current combined with the pixel size (about 5 nm of beam spot for 5 nm of pixel resolution). This last aspect confirms that to achieve good results in EBL, the size of the pixels in which the design is fractured has to be 10 times smaller than the critical dimension of the design. And then the beam current has to be chosen to best fit these dimensions. Then, also the dose plays a role, but if all the other parameters are optimized, the window of optimal doses is very large (in this case

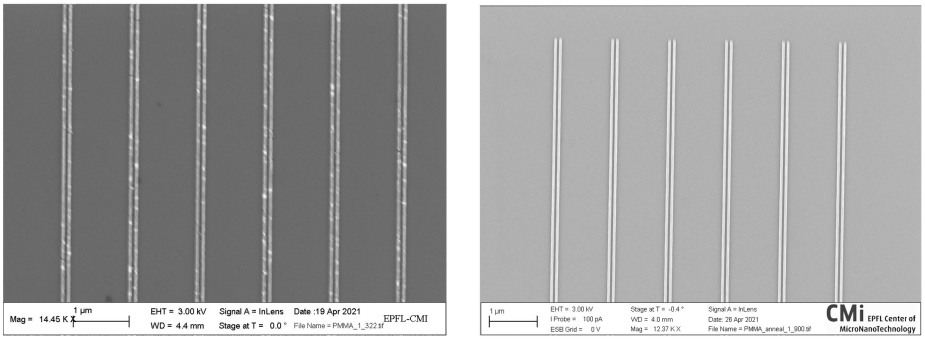
almost 1000 $\mu C/cm^2$). However, the dose plays anyhow a role: higher doses, as already anticipated, lead to larger features, as shown in Fig.3.25 Also the quality of



(a) **(b)**

Figure 3.25: SEM picture of two different couple of lines exposed respectively with 900 (a) and 1400 (b) $\mu C/cm^2$. As the dose increases, also the dimension of the patterned lines does, thus reducing the gap between them.

the patterned metal is quite good and the few grains visible disappear after the 15 minutes RTA at 400°C. Fig.3.26 reveals this aspect, confirming the positive aspects of the RTA also in improving the conformity of the metal gates. The just obtained



(a) **(b)**

Figure 3.26: SEM picture of the array of Ti-Pd lines before (a) and after (b) the RTA. The clearly visible grains disappear after the thermal treatment, this improving the conformity of the metal.

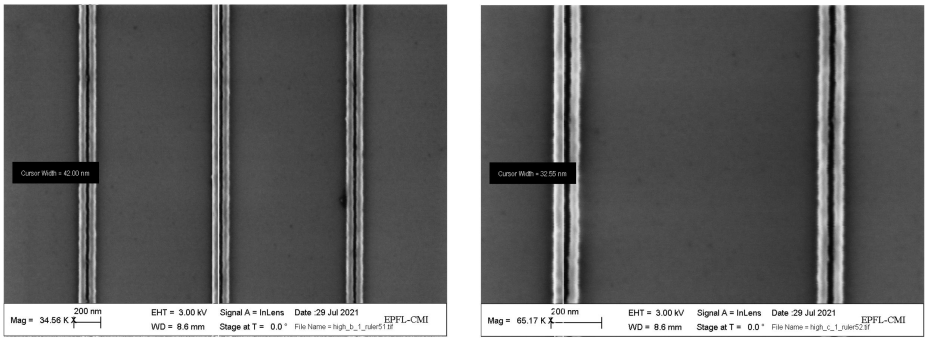
results are now finally acceptable, but since the yield in this case is almost perfect, the next step is to try to push toward even smaller features, to understand the real limits with the tool at disposal, at least with PMMA, a very easy-to-treat resist with respect to other like CSAR or HSQ, which are also more expensive.

4) Heavy PMMA for features smaller than 50 nm.

The last resolution test consists in trying to pattern lines with the following dimensions, :

1. 40 nm gates separated by 40 nm gap;
2. 40 nm gates separated by 30 nm gap;
3. 30 nm gates separated by 30 nm gap;
4. 30 nm gates separated by 20 nm gap;
5. 20 nm gates separated by 20 nm gap.

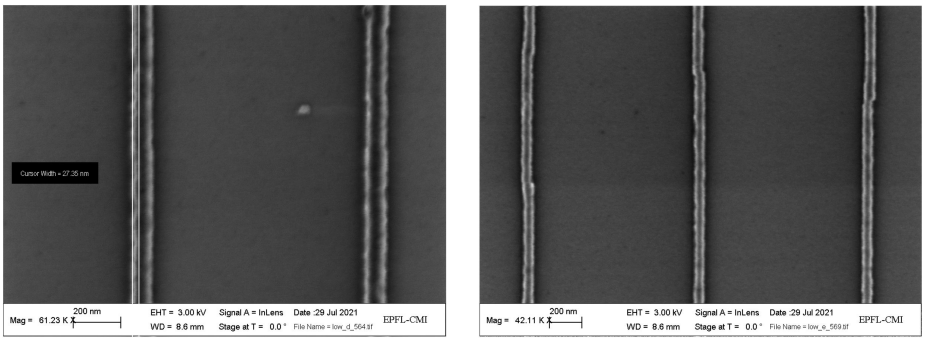
Two different currents are used: 2 and 0.7 nA, corresponding to a beam size of 5.1 and 4.5 nm, whereas the doses go from 700 to 1150 $\mu C/cm^2$, with a step of 50 $\mu C/cm^2$. The results show that the results are very good in the first three cases for both currents and all the doses (see Fig.3.27), whereas when the gap becomes less than 30 nm (i.e. 20 nm) the yield drops dramatically and only very few combinations show acceptable results, even though the quality of the metal is low and also the quality of the SEM image does not allow to fully understand the result obtained, as shown in Fig.3.28. In conclusion, under the light of these



(a) **(b)**

Figure 3.27: SEM image of the 40 nm gates separated by 30 nm of gap (a) and 30 nm gates separated by 30 nm (b). The beam current is 2 nA for these pictures, but no appreciable differences are noticed for the other smaller current (0.7 nA).

results, it can be safely said that the technology at disposal allows to pattern 30 nm gates separated by 30 nm of gap, whereas for lower gates and gap widths the results are not fully satisfactory. As a consequence, for the realization of SET, devices with three different dimensions of the barrier gates will be realized: 40 nm gates separated by 30 nm, 30 nm separated by 30 nm and 30 nm separated



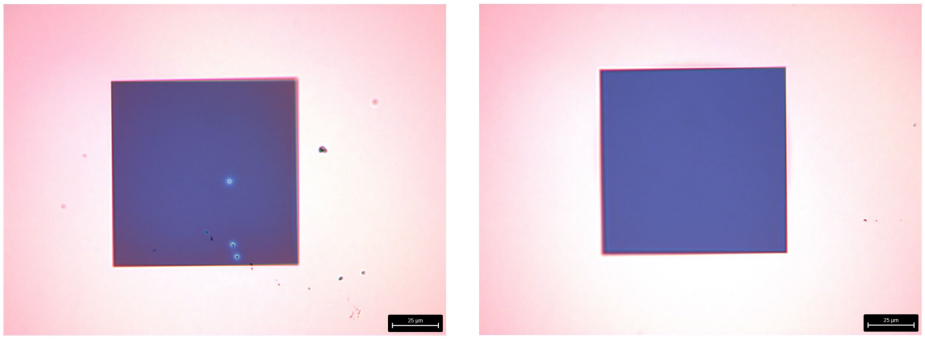
(a) **(b)**

Figure 3.28: SEM picture of the 30 nm gates separated by 20 nm of gap (a) and 20 nm gates separated by 20 nm (b). The quality of the image is not so good, but the features seem to be patterned correctly, even though the quality of lift-off is very low. Both the two images are referred the the smallest current utilized, i.e. 0.7 nA.

by 20 nm, being this last one the first border line case. The top gate will be designed to have a 40 nm width. Indeed, if the barrier gates are patterned correctly after metallization and lift-off, the same process repeated for the top gate will automatically work also for the top gate. That is because when features very close to each other are exposed, some *proximity effects*, due mainly to backscattered and secondary emitted electrons, can expose the resist also in the unwanted regions [64]. Concerning the dose, for the SET the lowest tested one of 700 $\mu C/cm^2$ is chosen: in this way the resist is fully exposed, without increasing the width of the features.

3.2.2 HSQ as Bond Pad Protection

The same test has to be repeated also for the other EBL resist used: HSQ. The features to pattern with this resist are very large (100 $\mu m \times 100\ \mu m$, the one of the metallic pads) and therefore a current-dose test is not needed. In fact, for features of hundreds of μm, the largest available current is used (200 nA, which gives a spot of 101.8 nm), together with the larger pixel resolution, i.e. 100 nm. The two parameters to find are the dose, since it can vary from 700 up to 7000 $\mu C/cm^2$, and the developing time, since it can vary from 30 seconds to 2 minutes and can be done both in TMAH (*Tetramethylammonium hydroxide*) 25% or in MF-CD-26, a TMAH based developer with a lower concentration of this latter. Doses from 700 to 1200 $\mu C/cm^2$ are tested, as well as an increasing developing time from 30 seconds to 2 minutes in MF-CD-26 developer. Fig.3.29 shows the result of the two extreme cases in the dose test: no evident differences are clearly visible, but the



(a) (b)

Figure 3.29: Optical microscope picture of the HSQ negative tone resist, exposed with a dose of 700 (a) and 1200 $\mu C/cm^2$ (b). The second one shows sharper edges, indicating a higher crosslink after exposure. Both squares are developed 2 minutes in MF-CD-26.

one with the highest dose has sharper edges. The developing time for both is 2 minutes and their thickness is 120 nm, sufficient to ensure protection of the below gate oxide from wire bonding.

3.2.3 Alignment Tests

The next problem to face when dealing with lithography is the alignment of subsequent layers on the previous ones. If the features to pattern are large, misalignment errors usually are not detrimental for the device, but when they start to be in the order of the nm, alignment becomes a critical step in device fabrication. The used technique for alignment consists in pattering metallic markers, which are found by the tool and used as reference align for subsequent layers.

Alignment with MLA150 Direct Laser Writer

As already illustrated in the process flow, MLA150 is used to open the window above the dot where the thick Al_2O_3 has to be etched (to reduce the charge noise), to re-open the contacts after the deposition of the oxide passivation layer which packages the finished device. Each of these two lithographic step needs to be aligned onto the previous layer and, for this reason, the same square Ti/Pt markers of 20 μm × 20 μm used for EBL alignment are used also for the MLA150 tool. The software knows where the markers are, according to the layout file (*.gds*), so it moves the stage to the first one, then a camera allows to visualize the marker on a screen and the operator has to collocate the cross visualized on the screen into the

center of the marker. The alignment is therefore done manually with the help of the software, since it is the user who aligns the cross onto the already patterned markers. Once the software knows where the four markers are physically located, since it knows also their positions in the layout, it uses them to find the centre of the design. At least three markers, in three different corners of the chip, are necessary to correct its offset and rotation.
MLA150, from its specifications, has a maximum misalignment error of 500 nm, fully acceptable for the purpose for which it is used in the SET fabrication (the window to etch is 20 μm × 20 μm and the metal pads to be re-opened are 100 μm × 100 μm). In fact, as shown in Fig.3.30, the contacts are re-opened (thanks to BHF) in the correct position after having deposited a SiO_2 layer above the first Ti/Pd of some MOS capacitors used as test structures.

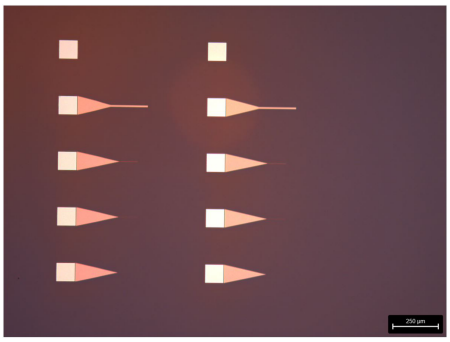


Figure 3.30: Re-opened contacts perfectly aligned on the already deposited Ti/Pd MOS capacitors.

Alignment with E-beam Lithography

A little bit more complicated is the alignment procedure for EBL. In fact, this alignment, which can be automatic or manual, is based on the same just illustrated working principle: markers are used to find the new centre of the layer to print, adjusting offset and rotation. However, in EBL other factors play an important role, like thermal expansion and contraction of the substrate and beam drift during exposure. For this reason, usually more than one set of markers is used: fixed global markers at wafer level, searched by the tool only once at the beginning, and then local markers in each chip. This approach limits the surface variations along the surface, which play a role when the target precision in the alignment is of few nm. If the automatic alignment is chosen, the software is able to find by itself the markers in a range of few decades of μm , as long as the contrast between the metal and the substrate is good. Only in this way, in fact, the tool is able to identify the exact edges of the markers, as shown in Fig.3.31. Since the field

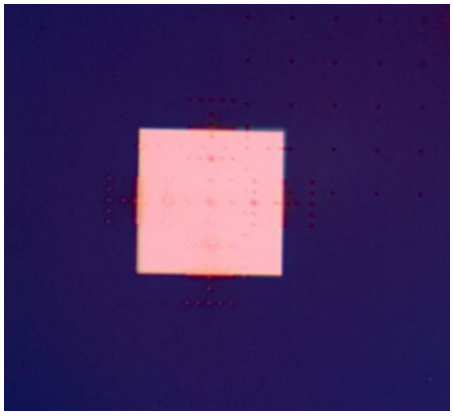
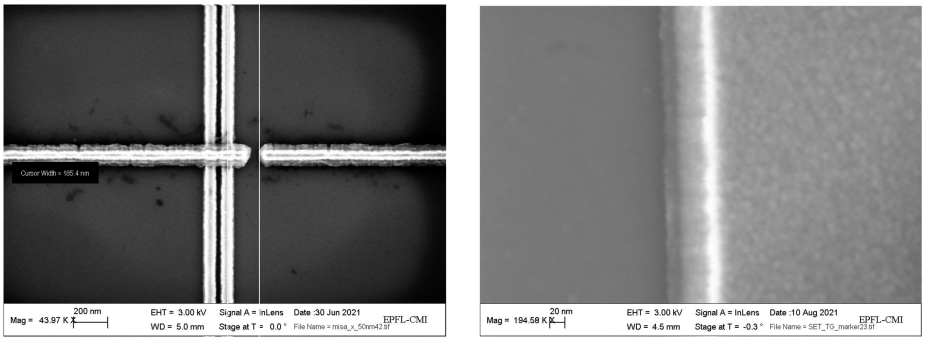


Figure 3.31: Ti/Pt marker found automatically by the e-beam tool.

of view of the tool when looking for the markers is limited, the position of the first marker (called *pre-marker*) has to be provided with accuracy. To do so, an alignment microscope allows to measure the exact position of this marker from a reference point called *Faraday cup*, which is the same reference used by the tool. Then, a manual alignment, has to be performed by the user so that when the tool will look for the second marker, it is able to find it at the position indicated. If, for any reason, the tool is not able to find even one of them, a manual alignment is necessary: the SEM is switched on the procedure becomes exactly the same of the one for the MLA150 laser writer. Usually, the automatic alignment, since done by the software, is more accurate than the manual one, so the first one is preferred in the context of this work. The results of the first alignment test are shown in



(a) **(b)**

Figure 3.32: SEM images of the misalignment error measured after the first test (a) and a Ti/Pt marker patterned with a pixel resolution of 5 nm (b). The marker shows an edge roughness of not more than very few nm.

Fig.3.32a. A misalignment of about 180 nm is observed on the x-direction, since the horizontal gates, similar to the lead gates of the designs proposed, in Chapter 2.1, which were supposed to land exactly on the middle of the two vertical barrier gates, are clearly shifted. Initially, this misalignment error was attributed to the non-optimal quality of the markers. In fact, this first set of markers was realized in Ti/Pd and not in Ti/Pt and with a resolution of 10 nm and not 5 nm like all the subsequent ones (see Fig.3.32b). Marker's edge roughness can introduce a misalignment of some nm if the edge is not perfectly sharp, but it does not explain a so huge misalignment (180 nm for EBL are too much). This is the reason why, not understanding initially the reason for this misalignment, the design of the SET was simplified to just two gate layers: the barrier gates and a top gate, without splitting the role of the top gate in two different layers to have a gate dedicated just to the electron occupancy inside the dot control (plunger and lead gates to bring electrons close to the dot from the ohmic regions). In fact, being the QD used just as a SET and not a spin qubit (where the number of the electrons has to be between 1 and 4), this design change does not affect the performance of the device and allows to better stand misalignments of this order of magnitude, if the layout is properly thought.

The explanation is finally individuated, in a problem in the conversion from the *.gds* file into the *.gpf* before and into the *.cjob* then. In fact, the software CATS does not center the design in the center of the *.gds* layout file, but according to the limits set when importing the different layers. Sometimes, if the limits are not always the same when loading the new layers to be printed, the centers of the different layouts can be different, thus leading to a misalignment error. The problem was solved switching to the BEAMER software.

3.2.4 Etching Tests

The last series of fabrication tests consists in wet etching tests in BHF and HF 1% to extract the etch rate and use it to correctly thin down the Soitec SOI wafer from 133 nm of gate oxide to only 5 nm. These test are carried out on the backside of the SOI wafers from Soitec (three in total, but only two have been used for this purpose), after having protected all the frontside with a thick photoresist (1.1 μm of AZ 1512 HS). For this kind of test, the key tool used is the *ellipsometer*, an optical instrument used to extract parameters like the thickness of a thin film or its refractive index thanks to property of light to change its polarization after the interaction with the sample. In fact, light can be reflected, transmitted or absorbed after the interaction with the sample and so, if the polarization is known before and after the interaction, parameters like the thickness and the refractive index of the sample can be extracted [65]. All the measurements are taken in five different points of the wafer, labelled with letters from A to E in the following tables.

First BHF test on SOI backside

The first test consists in a wet etching in BHF (50% HF/40% NH_4F) of the backside of one of the three SOI wafers. The expected reported etch rate of a dry thermal oxide in BHF is of 86 nm/min, so a first etching cycle of 45 seconds is performed in order to etch about 65 nm of oxide, so to halve its thickness. The results of the ellipsometer measurements before and after this etching step are reported in Table 3.1: The first results show an homogeneous etch rate, even tough

	Before etching (nm)	After etching (nm)	Etch rate (nm/min)
A	133.378	72.129	81.67
B	133.885	72.176	82.28
C	133.714	72.129	82.11
D	133.378	73.258	80.16
E	133.375	72.114	83.01

Table 3.1: Ellipsometer measurements of the SOI wafer on the backside before etching, after etching in BHF for 45 s and the corresponding etch rate.

the etch rate is a little bit lower than expected (81.85 nm/min in average). With the extracted etch rate, a second etching cycle is done in order to etch 52 nm and leave only 20 nm of oxide. The computed time is of 38 s and the results are shown in Table 3.2: This time, the etch rate is higher than the previous case (88.85

	Before etching (nm)	After etching (nm)	Etch rate (nm/min)
A	72.129	15.533	89.36
B	72.176	15.397	89.65
C	72.129	14.419	91.12
D	73.258	18.112	87.07
E	72.114	16.988	87.04

Table 3.2: Ellipsometer measurements of the SOI wafer on the backside before and after the second etching step in BHF of 38 s, with the corresponding etch rate.

nm/min in average) and, in fact, the result is about 15 nm of oxide instead of the predicted 20 nm. So, taking in consideration this last extracted etch rate, which represents the worst case, a third etching test in BHF is carried out to thin it down to 5 nm. The computed etching time is 6 s to etch about 9 nm of $SiO2$ leaving still a margin of some nm. However, the results, reported in Table 3.4, show that the oxide is completed removed in almost all the five points, where a very thin native

oxide of some $\mathring{A}$ has already re-grown. These results highlight the fact that wet

	Before etching (nm)	After etching (nm)
A	15.533	0.299
B	15.397	0.462
C	14.419	0.295
D	18.112	0.722
E	16.988	0.195

Table 3.3: Ellipsometer measurements of the SOI wafer on the backside before and after the third (and final) etching step of 6 s in BHF.

etching, being led by a chemical reaction, is characterized by a reliable etch rate only if the procedure is long enough to let the reaction stabilize. However, if the etching time is too short, the reaction does not have enough time to stabilize and the initial etching rate is higher than expected.

HF 1% test for a more controlled etching

To overcome this problem, a smart solution consists in splitting the etching procedure in two different etchings: the first one in BHF to etch the thicker part and then a second one in HF 1% , with a definitely lower (and so more controllable) etch rate. So, a second test on the backside of a second SOI wafer is performed. This test demonstrated the reliability of the two etching steps in BHF to etch most of the oxide (45 s to etch from 130 nm to 70 nm and other 33 s, instead of 38 s, to go down to 20 nm, instead of the 15 nm of the first test) and allowed to extract the etch rate for the 1% diluted HF: 3 nm/min. This result is lower than the expected theoretical one (5 nm/min), but was confirmed by three gradual etching steps and so used also in the final etching test on the frontside of one SOI wafer. For this last test, again a preliminary etching in BHF (45 s and then other 33 s) allows to thin down the oxide to 20 nm. Then, three cycles of etching in HF 1% and ellipsometer measurements produced the results reported in Table 3.4. These results confirm the etching rate of HF 1% being of about 3 nm/min (2.95 nm/min, 3.12 nm/min and 2.82 nm/min after the first, the second and the third etching step respectively) and allow to reach a final gate oxide thickness between 5 and 7 nm. This result is acceptable for the SET realization, but the variability across the wafer forces to take an average thickness of 6 nm of gate oxide to do the TCAD simulations for the implantation, as it will be explained in the following chapter. Luckily, the second batch of SOI wafers from CMi, as already explained, having enough top Si, allow the complete etching of the thick oxide in BHF (the one grown to thin down the top Si) and since this etching is selective, i.e. silicon is just very slowly

	BHF, 45 + 33 s (nm)	HF 1%, 1 min (nm)	HF 1%, 2 min 40 s (nm)	HF 1%, 1 min 20 s (nm)
A	20.598	17.508	9.391	5.312
B	21.176	18.439	9.373	5.164
C	20.946	18.777	10.521	6.251
D	21.424	18.272	10.248	7.342
E	21434	17.810	9.572	6.213

Table 3.4: Ellipsometer measurements of the SOI oxide on the frontside after the first two etching in BHF (45 + 33 s, first column) and after the three etching steps in HF 1%. The etch rate for HF 1% is about 3 nm/min after each intermediate step.

attacked by BHF with respect to the amorphous oxide, these etching test are not needed. For this batch, the oxide is completely removed a thin precise gate oxide of 5 nm is grown separately with a second oxidation process.

Chapter 4

SOI SET Layout and TCAD Simulations

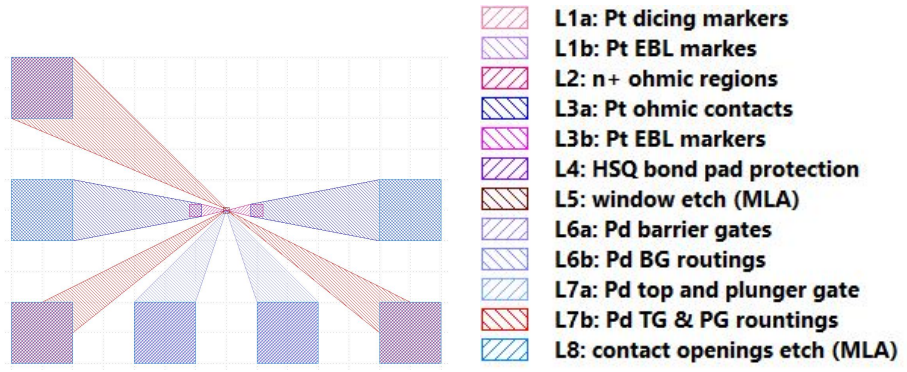
Under the light of the considerations made so far, the full design of the device can be finally described. The aim of this chapter is to present the detailed design of the SET, which basically means illustrating the different layers of the layout printed during the different e-beam and photolithographies. Moreover, being part of the design preliminary to device fabrication, in this chapter are also presented the results of the TCAD simulations, used to find the best parameters (i.e. energy and dose) for the doping of the source and drain ohmic regions.

4.1 SET Layout

The SET design is realized through the software KLayout, which allows to easily draw the desired shapes, differentiating them according to the layer of the layout they belong to. It also allows to organize the design into cells, which can be repeated inside the full layout (for example a chip repeated several times in the wafer). In this section, three different hierarchies are illustrated with a *bottom-up* approach, which means from device level to chip level (since more devices with different designs are present on the same chip, as well as test structures for different purposes) and from chip level to wafer level (since all the steps until implantation are realized at wafer level).

4.1.1 Device Design

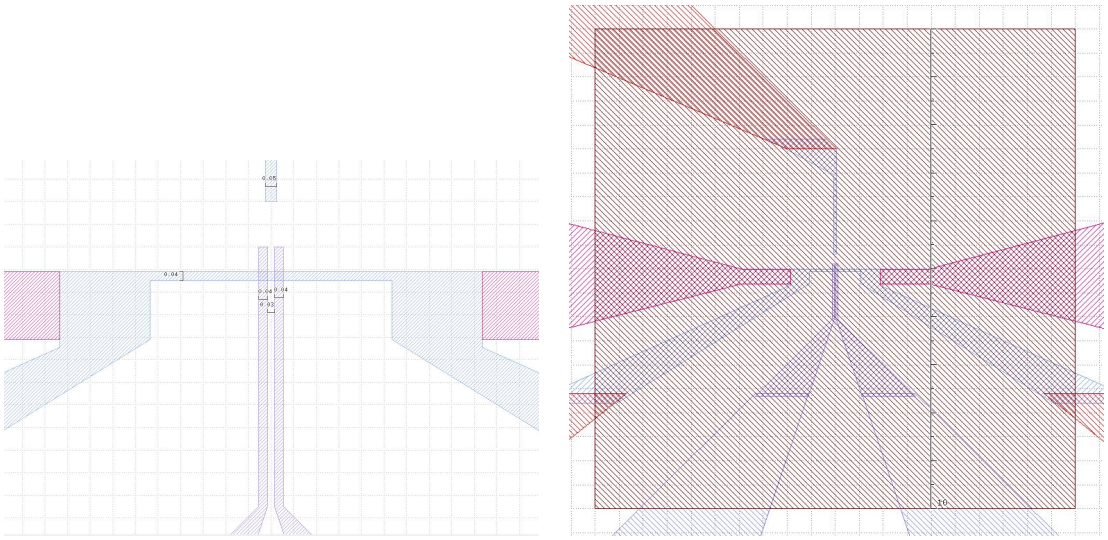
First, the layout of the SET is presented, constituted not only by the active region where the QD is defined, but also by the long routings to connect the gates and the ohmic contacts to the big metallic pads, as shown in Fig.5.5a. In Fig.4.1 are



(a) **(b)**

Figure 4.1: Overview of the SET routings and pads (a), with the corresponding legend showing all the layers involved in the layout, except from the first layer (L1) and part of the third one (L3b) which include the markers present at chip level. Metal pads for the contacts are $100\mu m \times 100\ \mu m$, whereas the ohmic contacts $20\mu m \times 20\ \mu m$.

reported the different layers of the layout. Even though twelve different layers are reported, the number of lithographic steps is actually eight, two of which, as already explained in the process flow, are realized with standard photolithography (laser writer MLA150), i.e. the window etching in the dot region and the contacts re-opening. The reason why twelve layers are present is to exploit the possibility of printing the same layer with two different currents with e-beam lithography during the same exposure. In fact, the fine gate structures, as well as the high quality Pt markers for EBL alignment, are patterned with a pixel size of 5 nm, but the larger structures of the same metallic layer, like the gate routings, the dicing markers and the ohmic contacts cannot be fractured with pixels of the same size, otherwise the exposure would really last hours and hours for each layer. As a consequence, a smart solution consists in splitting the exposure into two different layers, so that features with different size can be patterned with two different beams. When doing so, in the case of the Ti/Pd fine gates with the long routings, a little overlap between them is necessary already in the layout, so that even if there are some drifts in the beam from one exposure to the other, their are still joined together and not separated. A smarter approach to solve this problem consists in forcing the tool looking for the markers not only before the first exposure, but also before the second one (for example when switching from the long routing to the fine gates). Further considerations about this aspect will be illustrated in the next chapter, the one dedicated to the fabrication results. This overlapping between the fine gates and the routings is visible in Fig.4.2b, the one where also the window of $10\mu m \times$



(a) **(b)**

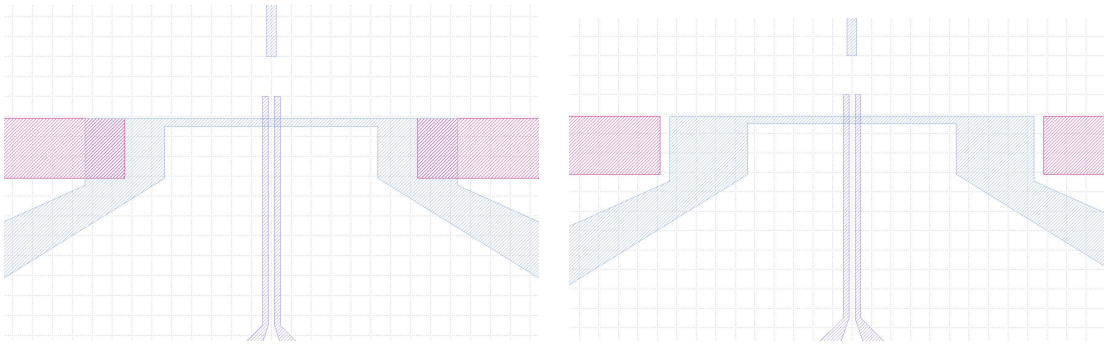
Figure 4.2: SET active region. a) Zoom on the QD region defined by 40 nm-width barrier gates and 50 nm-width top gate. b) $10\mu m \times 10\ \mu m$ window to be etched above the SET active region.

10 μm above the dot is. As already explained, the thick Al_2O_3 deposited by ALD to avoid leakages from the long routings (5 nm of thin SiO_2 are not sufficient) is etched away in the dot region to reduce the charge noise due to the traps in the thick oxide and at the SiO_2 and Al_2O3 interface. Fig.4.2a, on the other hand, is a zoom on the QD region, defined by a top gate, which creates the conductive channel, and the two barrier gates, which locally deplete it, creating two tunneling barriers which separate the electrons inside the dot from the ones in the channel. In front of the dot, a plunger gate is used to characterize the sensitivity of the dot (see Fig.2.6b). In fact, it can accumulate electrons, inducing a change in the electrostatic environment sensed by the SET. The SET is concluded by the two n+ ohmic regions, which act as source and drain, heavily doped to realize ohmic contacts used to measure the current.

4.1.2 Chip Design

Fig.4.2a of the last paragraph shows 40 nm barrier gates separated by a gap of 30 nm. However, EBL resolution tests demonstrated that also 30 nm gates separated by 30 nm-gap are possible to realize, whereas gates with same widths, but separated by just 20 nm represent the border line case. For this reason, three different batches of SET are present on the same chip: the first one with 40-30-40 nm gates, the

second with 30-30-30 nm ones and the third with 30-20-20 nm gates. Each SET is repeated nine times in the same row, so that device variability can also be tested. A further degree of freedom is represented by the shape and the position of the ohmic implanted regions. In fact, achieve a good contact on UTB SOI wafers can be very challenging because of the high access resistance provided by ultra-thin top Si. To limit this problem, large contact areas between the heavily doped silicon and the metal can be realized and this is done in a further batch of SET on the chip. Moreover, since the diffusion of dopants during the activation annealing in a UTB SOI is a phenomenon that cannot be simulated with 100 % of reliability with TCAD simulators, as well as the implantation actually, to be safer other two batches of SET are designed so that, in the first, the ohmic regions are a little overlapped with the top gate, whereas in the second they are a little bit separated, as shown in Fig.4.3 In fact, in the first case, even in case of little misalignment, a larger



(a) **(b)**

Figure 4.3: Alternative design for the ohmic regions. a) Ohmic regions are overlapped by 200 nm to ensure a larger low-resistance area between the n+ region and the active channel. b) Gap of 50 nm between the n+ region and the channel to stay safe from uncontrolled dopants diffusion and/or misalignment errors.

low-resistance region between the channel and the ohmic region is ensured, whereas in the second, even in case of larger dopants diffusion after the activation annealing (or, also in this case, of little misalignments) the dopants stay at safety distance from the dot, whose presence would really be detrimental for the device (especially for the charge noise). In conclusion, different batches of SET, replicated nine times in the same row, are present, including different barrier gates dimensions and separations, different ohmic regions dimensions and positions, to cover more or less all the possible cases. These cases are all represented in the chip overview in Fig.4.4, where also some test structures like snake-shape metal lines, MOS capacitors and source and drain of MOSFETs are patterned to extract some useful parameters (like doping, resistivity, threshold voltage, traps and, eventually, channel mobility),

and whose zoom is provided in Fig.4.5. In Fig.4.4 are also shown the three sets of

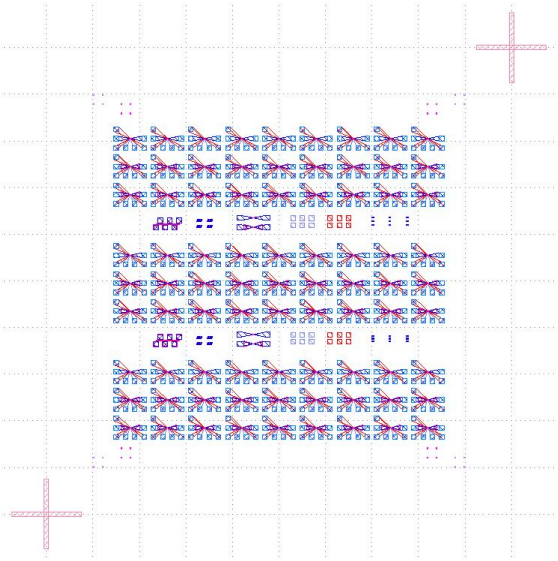
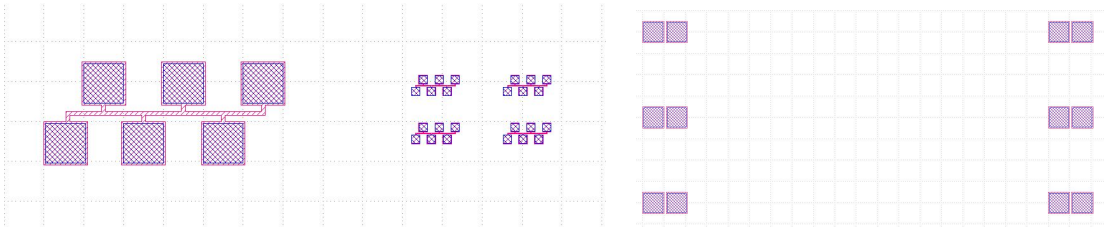


Figure 4.4: Full chip layout, including all the different variants of SET devices, as well as test structures for parameter extraction and the three sets of markers, one for the dicing and two for the EBL alignment.



(a) **(b)**

Figure 4.5: Test structures for parameters extraction. a) Snake-shape metal lines for doping and resistance of the ohmic contacts extraction. b) MOSFETs source and drain. The gate is not present at this stage and the channel has the same length of the SET one (2 μm).

markers used: the big crosses which defines the borders in which chips have to be diced and the two sets of EBL markers, which are patterned twice since the first set are damaged by the high temperature reached during the dopants activation annealing (900°C).

4.1.3 Wafer Design

All the steps subsequent to implantation are realized at chip level. However, the two preliminary lithographic steps of the process flow, i.e. the first markers set (for dicing and EBL alignment, shown in Fig.4.6) and the regions to be implanted, are patterned at wafer level, being the implantation a service offered by an external fab. As shown in Fig.4.7, not the whole wafer is patterned, but only a region equivalent to 15 chips, in order to use again the very expensive SOI wafer for other devices or to have some spare of wafer in case the implantation is not successfully.

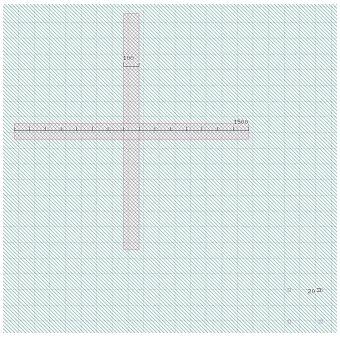


Figure 4.6: Markers patterned during the first ELB step. a) Dicing markers consist in crosses of $1500\mu m \times 100\ \mu m$. b) EBL markers are $20\mu m \times 20\ \mu m$ squares.

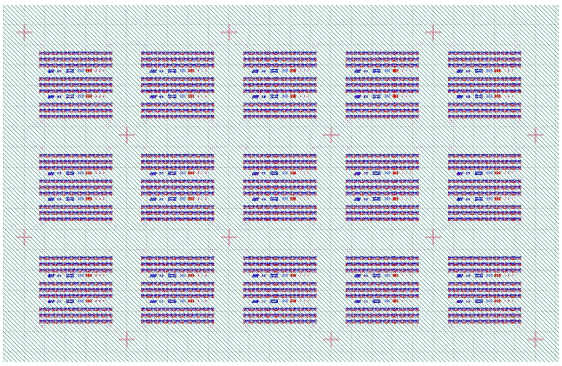


Figure 4.7: Wafer level layout showing the 15 chips. Only two lithographic steps are done at wafer level: markers and ohmic regions patterning.

4.2 TCAD Simulations for Ion Implantation

Doping by implantation requires the knowledge of both ions energy (in keV) and dose (in cm^{-2}). To get these data, preliminary TCAD simulations are performed through the aid of Sentaurus by Synopsis ®. The software allows to create the 2D geometry of the structure to simulate (the two different SOI wafers in this case) and, after defining the dimensions of the mesh, the model of interest is numerically solved in every element of the mesh. For ion implantation, Sentaurus employs a Markov Chain algorithm to describe the interaction between ions and the matter [66]. To include this model, two simple lines have to be added to the script:

- pdbSet ImplantData Montecarlo 1;
- pdbSet MCImplant model sentaurus.mc.

Simulations are optimized to achieve two essential goals:

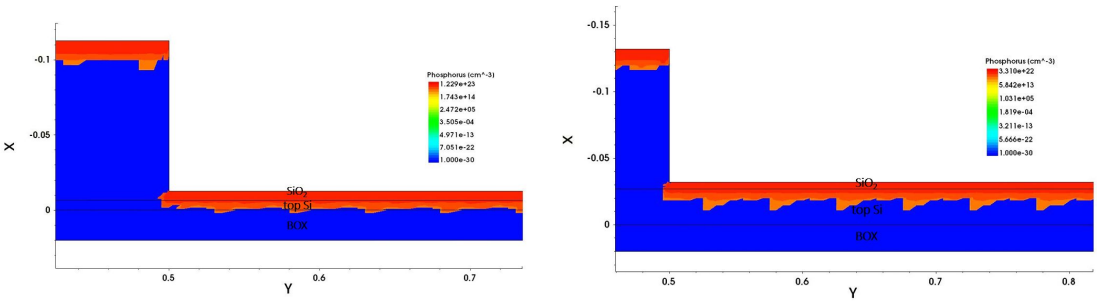
1. The BOX should not be damaged by the implanted ions;
2. The dopants concentration in the top Si after the activation annealing should be $\geq 10^{19}$ cm^{-3} to obtain a tunneling ohmic contact (see Appendix B.2). Results illustrated in Chapter 6.3.2 highlight that, to achieve a very low resistance, this values should be $\geq 10^{20}$ cm^{-3}. As a consequence this value is taken as target for the simulations.

The tilt angle is kept fixed to the standard one (7°) and phosphorous is chosen as donor dopant. In order to get the peak of the gaussian profile of the implanted dopants close to the surface, a 5 nm gate oxide is kept also for the 27 nm-thick top Si SOI wafers from CMi. On the contrary, for the 6/7 nm-thick SOI wafer from Soitec, the gate oxide has to be kept above the implanting regions, otherwise further top Si would be consumed by the re-growth of the native oxide. According to the ellipsometry measurements of Chapter 3.2.4, the gate oxide is $\approx 5/7$ nm thick, so a thickness of 6 nm is chosen for the simulation. The top Si thickness is fixed to 6.5 nm, being the measured one between 6 and 7 nm. The best parameters found for the two batches after the simulations are reported in Tab.4.1. Fig.4.8

	energy	dose
Soitec (6/7 nm)	0.9 keV	3×10^{16} cm^{-2}
CMi (27 nm)	1 keV	1×10^{16} cm^{-2}

Table 4.1: Best implantation parameters extracted from the TCAD simulations.

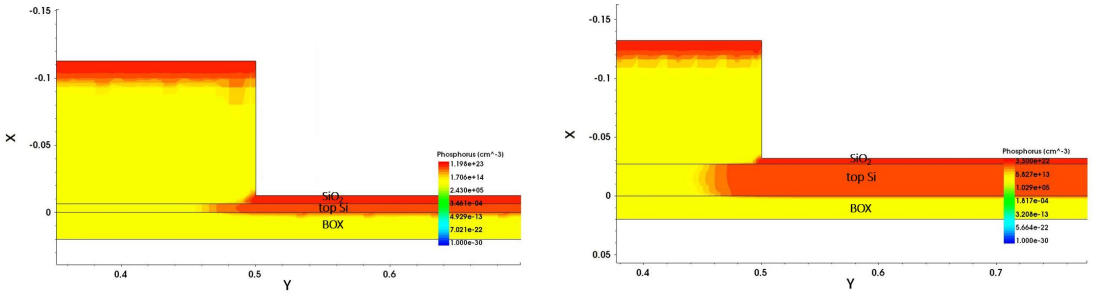
graphically shows the dopants concentration as implanted. It can be noticed that



(a) **(b)**

Figure 4.8: As implanted phosphorous concentrations corresponding to the parameters reported in Table 4.1 in the Soitec (a) and CMi (b) SOI wafers.

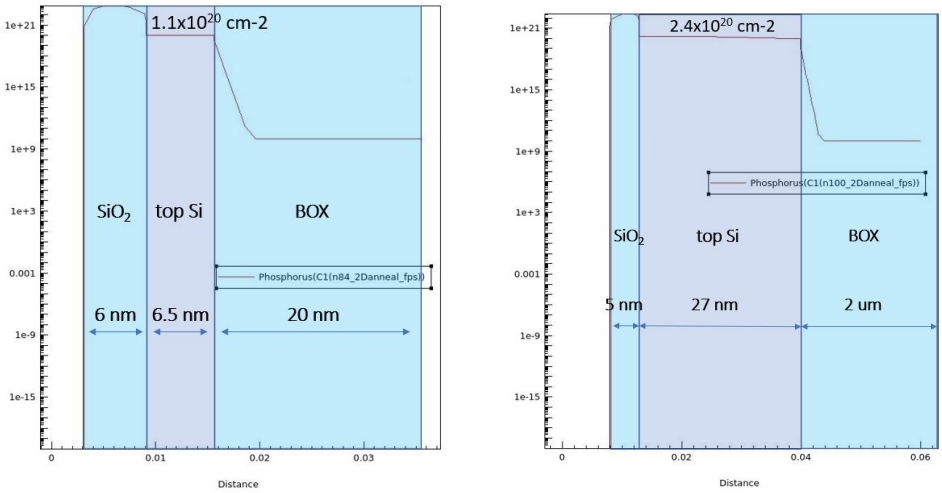
the chosen values of energy and dose allow to implant through the thin SiO_2 mask into the top Si layer, without damaging the BOX. This is completely true for the thicker SOI wafers, whereas a very little damage of the BOX can be observed in the thinner one. However, this latter cannot be prevented, because of the very thin top Si layer, which has more or less the same thickness of the oxide mask. Fig.4.9 shows the phosphorous concentration after the activation annealing step. A 5 s RTA at



(a) **(b)**

Figure 4.9: Phosphorous concentrations in the Soitec (a) and CMi (b) SOI wafers.

900°C is chosen for dopants activation, having demonstrated in this work that an high temperature RTA is detrimental for the oxide quality (see Chapter 6.3). For lower temperatures, dopants are not all activated, so 900°C has been individuated as the lowest temperature for this purpose. In order not to damage the oxide at all, a flash annealing should be performed. The two pictures highlight again that no dopants are present in the BOX, also after the annealing step. Also lateral diffusion is negligible (< 50 nm). The dopants profile after the annealing is obtained through a cutline along the X direction and it is reported in Fig.4.10 for both batches. It can



(a) **(b)**

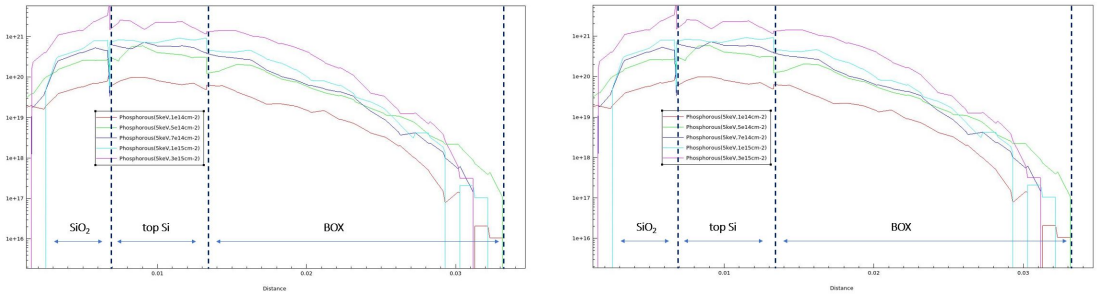
Figure 4.10: Dopants profile in the Soitec (a) and CMi (b) SOI wafers.

be noticed that, thanks to the both implant and annealing parameters optimization, the dopants concentration in the top Si layer is the desired one ($> 10^{20}$ cm^{-3}) and it rapidly decreases inside the oxide. Nevertheless, the company chosen for the ion implantation (*IBS*[1] in France), pointed out that the minimum energy for the beam line dedicated to ion implantation is 5 keV. In fact, for applications requiring very low energies and high doses such as FinFETs and DRAMs, doping by plasma immersion is used rather than ion implantation. However, this doping technique is very expensive and shows some drawbacks such as possible damage of the oxides and sputtering of atoms on the surface because of the plasma. As a consequence, first of all, simulations at 5 keV are performed to understand if a standard ion implantation at the lowest allowed energy was still feasible. Fig.4.11 shows the as implanted dopants profile at the energy of 5 keV and different doses. The figures confirm that doping cannot be carried out with these parameters. In fact, the phosphorous concentration remains high in the BOX for both the types of of SOI wafers and for all the simulated doses. The only alternative to explore is the plasma immersion doping, which can also be simulated in Sentaurus. To do so, a line has to be added into the code to take into account the interaction between the accelerated ions coming from the plasma and the substrate:

- pdbSet MCImplant PlasmaEnergyDistributionModel Burenkov,

from the name of the engineer who developed this model [67]. However, when

[1]https://www.ion-beam-services.com/foundry/



(a) **(b)**

Figure 4.11: Dopants profile in the Soitec (a) and CMi (b) SOI wafers as implanted at 5 keV and for different doses.

doping by plasma, another phenomenon has to be taken into account: the sputtering of atoms from the wafer surface due to the dopant ions accelerated from the plasma toward the substrate. To make this phenomenon neglibile with respect to dopants implantation in the wafer, energies and doses have to be set properly. A free of use software, SRIM-TRIM [2], was used to extract the dose above which sputtering starts to dominate over implantation. From the software, a maximum dose of 1.1×10^{16} cm^{-2} was extracted. For this reason, the dose for the CMi SOI wafer was chosen equal to this value, keeping the energy at 1 keV. Then, considering that:

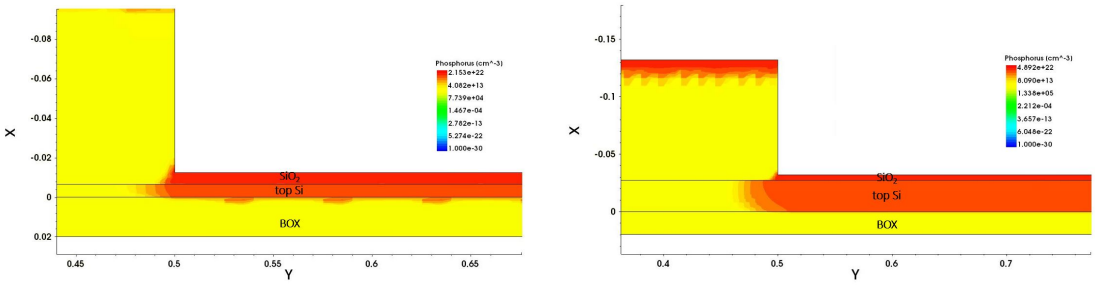
- this Burenkov model is very recent and only very few works have been published (and none with phosphorous). As a consequence the simulation results cannot be considered completely reliable;
- the Soitec SOI wafer has a very thin top Si film, which has to be protected from sputtering due to plasma;
- choosing two different energies for the two different batches of SOI wafers increases the cost of the service,

the energy and dose chosen for the Soitec SOI wafer are different from the previous values for standard ion implantation. The new values are reported in Table 4.2. The dose is lowered to avoid sputtering, whereas for the other thicker SOI wafer, even if some sputtering occurs due to simulations uncertainty, the wafer can still be used for devices realization and so the dose is kept the maximum one. Simulation results are shown in Fig.4.12 and Fig.4.13. The simulations show very conform doping profiles and minimum damage to the BOX. However, the maximum concentration of dopants is estimated to 3-4$\times10^{19}$ cm^{-3}.

[2]http://www.srim.org/

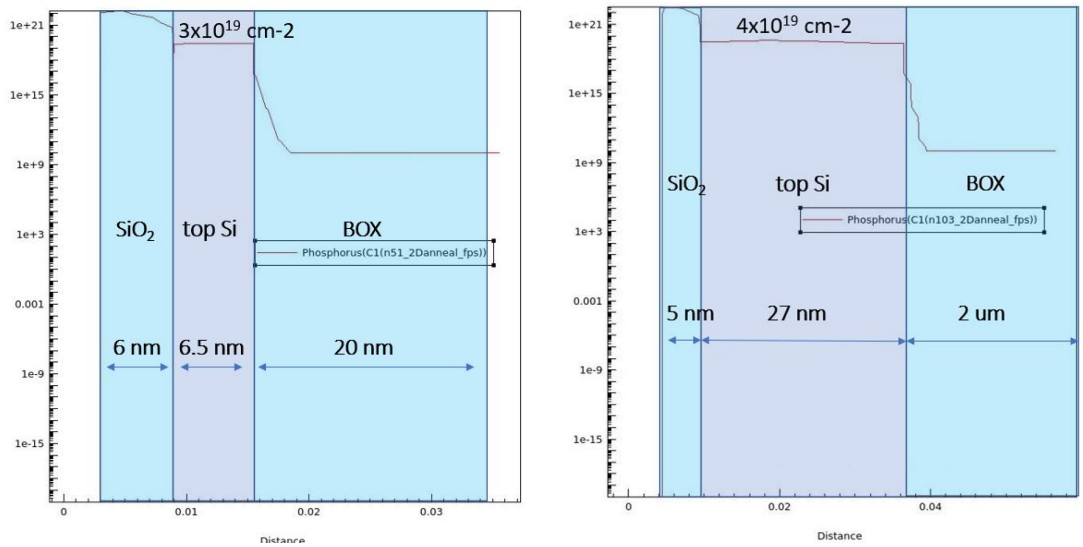
	energy	dose
Soitec (6/7 nm)	1 keV	5×10^{15} cm^{-2}
CMi (27 nm)	1 keV	1.1×10^{16} cm^{-2}

Table 4.2: Final parameters for plasma doping.



(a) **(b)**

Figure 4.12: Phosphorous concentrations after annealing for the Soitec (20 s at 900°C) (a) and CMi (5 s at 900°C) (b) SOI batches.



(a) **(b)**

Figure 4.13: Dopants profile after annealing for the Soitec (a) and CMi (b) SOI wafers.

Chapter 5

SOI SET Fabrication Results

This chapter presents the experimental fabrication results. Due to the long times necessary to do the preliminary tests, get the SOI wafers, find all the best parameters for plasma implantation, ship the wafers to the external fab and get them back, the fabrication of the SOI SET has not been finalized in the short time of this Thesis. However, the entire process flow illustrated in Chapter 3.1 has been validated. In fact, all the steps until the EBL to define the ohmic regions have been performed on the two batches of SOI wafers, whereas all the others following the implantation on dummy 1 cm $\times$ 1 cm silicon chips from test wafers. In this way, when the SOI wafers will be back from implantation, all the steps have already been previously tested and optimized and just need to be repeated onto SOI substrates. Moreover, this approach allows also not to waste expensive SOI chips, testing all the steps before on bulk Si. The only drawback of this approach consists in the fact that the dummy chips do not have the implanted ohmic regions and so the finalized dummy SET cannot be tested. For this reason, the characterization results shown in Chapter 6 are not related to the SET characterization (the stability diagrams, quantum dot spectroscopy and charge noise characterization explained in Chapter 2), but to the oxide quality and ohmic contacts optimization, carried out in parallel. The fabrication results of the MOScaps and the ohmic contacts used for these studies are reported in this chapter as well.

5.1 SOI SET (from Beginning to Implantation)

The two batches of SOI wafers consist in respectively three wafers from Soitec and two wafers from CMi. The first ones have an initial top Si thickness of 70 nm and ad BOX of just 20 nm. In the second batch from CMi, instead, the initial

top Si thickness is of 220 nm and the BOX is 2 μm thick. Table 5.1 shows the ellipsometry measurements of the five wafers at disposal in the central point. The

	Native Oxide (nm)	Top Si (nm)	BOX (nm)
Soitec 1	1.245	69.922	21.584
Soitec 2	1.265	69.858	21.398
Soitec 3	1.602	69.615	21.466
CMi 1	1.247	218.632	2011.358
CMi 2	1.215	218.965	2012.560

Table 5.1: Ellipsometer measurements of the five SOI wafers at disposal.

ellipsometer measurements are in good agree with the expected values, being also the maximum native oxide usually about 1.3 nm.

5.1.1 Thermal Oxidation for Top Si Thinning Down

In order to thin down the top Si and get a FD SOI wafer, a dry oxidation is performed to grow SiO_2 at the expenses of Si. To ask the precise thickness of oxide, first the new top Si thickness is measured just after the RCA cleaning, mandatory before oxidation and known to consume some Si. The ellipsometry results are reported in Table 5.2. For a better precision, necessary in this case, two different points of the wafers have been measured. RCA cleaning consumed about 4-5 nm

	Point 1			Point 2		
	Native Oxide (nm)	Top Si (nm)	BOX (nm)	Native Oxide (nm)	Top Si (nm)	BOX (nm)
Soitec 1	0.750	65.536	21.417	0.679	65.468	21.642
Soitec 2	0.729	64.818	21.396	0.729	65.135	21.496
Soitec 3	0.769	64.818	21.338	0.725	64.850	21.373
CMi 1	1.023	211.581	2011.274	1.024	211.275	2011.224
CMi 2	1.014	211.732	2012.320	1.026	211.796	2011.121

Table 5.2: Ellipsometer measurements of the five SOI wafers at disposal after RCA cleaning. Measurements are registered in two different points.

of top Si in Soitec wafers and even more in the second batch. Knowing the exact starting top Si thickness, it is possible to compute the oxide thickness to ask to the oxidation service in order to get the desired final top Si thickness (see Chapter 3.1 and 3.2). For the first batch, a 100 nm-thick oxide was asked, whereas for the second a 410 nm-thick one. These parameters, in fact, allow to reach the desired top Si thicknesses of 20 and 30 nm, respectively. Oxidation is performed at 850°C with DCE. However, the oxides obtained do not have the desired thickness, as

shown in Table 5.3, where all the measures are averaged on five points of each wafer. It can be immediately noticed that for the first batch, the oxide is 30 nm

	Grown Oxide (nm)	Top Si (nm)	BOX (nm)
Soitec 1	133.543	6.445	21.242
Soitec 2	132.644	6.120	21.104
Soitec 3	131.449	6.902	21.074
CMi 1	389.905	39.123	2014.043
CMi 2	393.198	37.453	2013.135

Table 5.3: Ellipsometer measurements of the five SOI wafers at disposal after the first oxidation.

thicker than expected, resulting in a remaining top Si of just 6-7 nm. This forces to use this oxide also as gate oxide after a proper thinning down (see Chapter 3.2.4), because no further Si can be consumed. For what concerns the second batch, the resulting oxide is slightly thinner than expected. For this reason, the top Si is still 37-39 nm thick instead of the desired 30 nm. However, for this second batch it is possible to etch away all the oxide and grow a new ultra-thin gate oxide.

5.1.2 Silicon Oxide Wet Etching

SiO_2 is then removed thanks to a BHF bath. For the first batch of UTB SOI wafers, to obtain the final desired thickness of gate oxide equal to 5 nm, a lot of etching tests have been carried out in both BHF and HF 1% baths (see Chapter 3.2.4). On the other hand, for the second batch, all the grown SiO_2 can be entirely removed in BHF. Table 5.4 shows the obtained results in five different points for the two processed wafers (one for each batch). The final gate oxide thickness for

	Soitec 1			CMi 1		
	Gate Oxide (nm)	Top Si (nm)	BOX (nm)	Native Oxide (nm)	Top Si (nm)	BOX (nm)
A	5.312	6.728	21.240	0.148	38.814	2016.418
B	5.164	6.331	21.186	0.106	39.128	2107.361
C	6.251	6.228	21.211	0.226	38.409	2016.350
D	7.342	6.483	21.231	0.159	38.050	2016.528
E	6.213	6.444	21.224	0.178	38.392	2013.994

Table 5.4: Ellipsometer measurements on five points of the two SOI processed wafers after the etching step.

the Soitec SOI wafer is between 5 and 7 nm, whereas the top Si is about 6 nm.

5.1.3 Gate Oxide Thermal Oxidation

This section concerns only the batch of thicker SOI wafers from CMi. In fact, after the etching of the thick oxide, a new 5 nm high-quality thermal oxide can be grown through thermal dry oxidation in DCE. Again the oxidation temperature is 850°C. The results in five different points of the CMi 1 SOI wafer are shown in Table 5.5. This time the gate oxide is perfectly uniform with respect to the first batch where

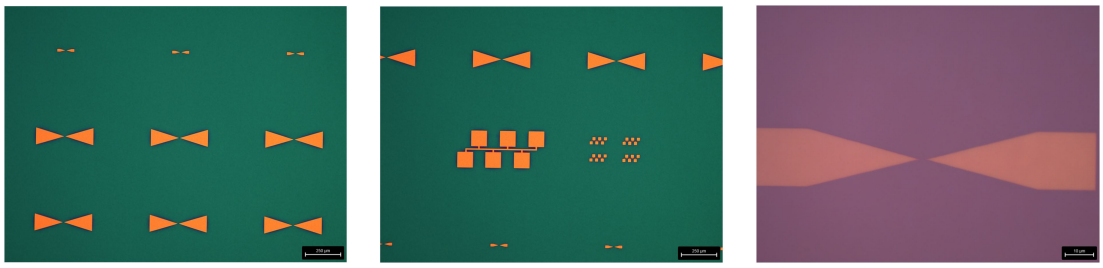
	Gate Oxide (nm)	Top Si (nm)	BOX (nm)
A	5.054	27.953	2019.273
B	5.065	28.341	2018.884
C	5.024	28.068	2018.625
E	5.135	27.462	2018.610
F	5.053	27.320	2018.421

Table 5.5: Ellipsometer measurements of one of the SOI wafers from CMi after the second oxidation.

the gate oxide is obtained by etching down the thick oxide. This constitutes a big advantage of the second batch with respect to the first, combined to the easier realization of ohmic contacts.

5.1.4 EBL for Ohmic Regions (n+ Source and Drain)

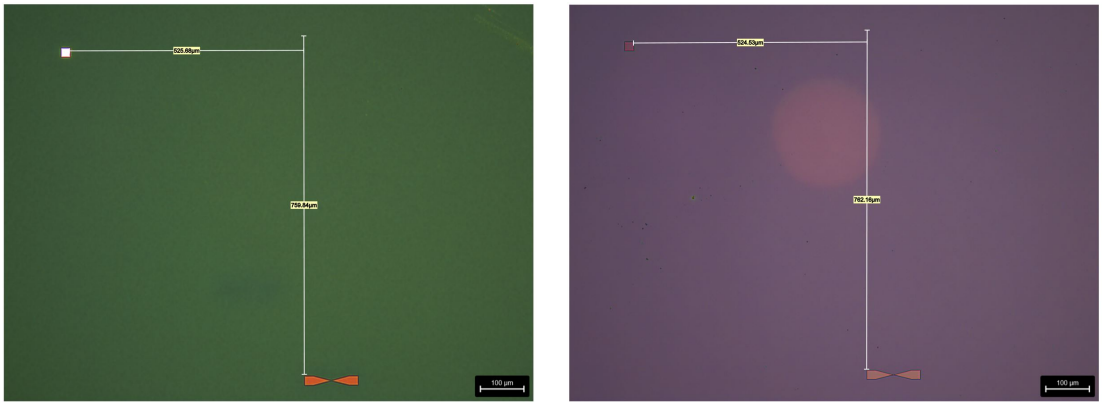
This is the second lithographic step after the very first one used to pattern the markers for dicing and for the alignment. A MMA/PMMA bilayer of respectively 200 and 90/100 nm is used as a mask to protect the non developed regions from dopants. The results, after EBL exposure and development, are shown in Fig.5.1.



(a) **(b)** **(c)**

Figure 5.1: Ohmic regions after exposure and development. a) and b) are from the thinner Soitec wafer, whereas c) from the thicker CMi one. The colour differences between the two batches is due to the different thicknesses.

Fig.5.2 shows the alignment of the developed resist on the Ti/Pt marker patterned during the previous lithographic step.



(a) **(b)**

Figure 5.2: Ohmic regions perfectly aligned to the metal marker. a) Soitec wafer. b) CMi wafer.

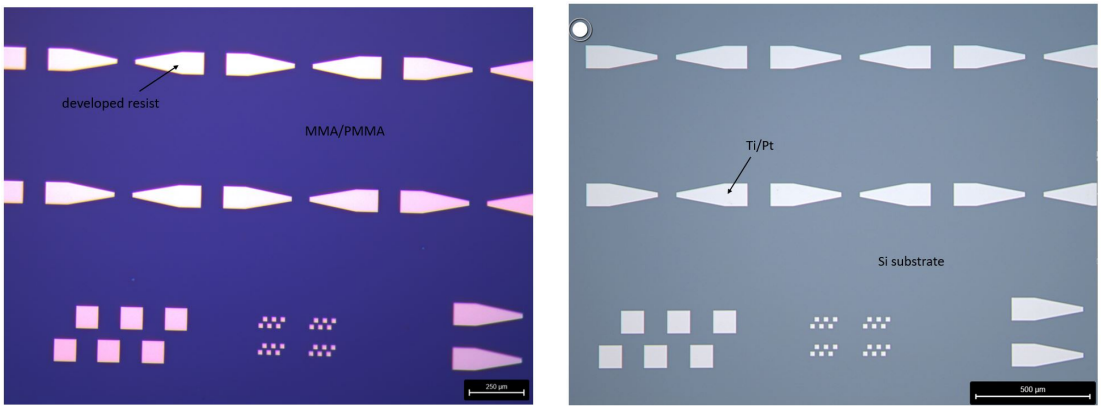
After this step, the SOI wafers are ready to be shipped to IBS in France for plasma implantation. As a consequence, all the following presented results concern the SET fabricated on a dummy bulk Si chip.

5.2 Dummy SET (from Implantation to the End)

The fabrication process moves now to the chip level. The aim of this dummy SET is to validate all the remaining steps of the process flow to fix all the eventual problems before starting processing the expensive SOI implanted wafers. For this purpose, a Si chip with 5 nm of gate oxide (coming from one of the wafers used for the MOS capacitors fabrication) is used.

5.2.1 Platinum Ohmic Contacts and Markers

The first step is the patterning and deposition of the Ti/Pt ohmic contacts for source and drain and markers for alignment. After EBL exposure and development, a quick etching step in HF 1% to remove the thin gate oxide in the contact (and markers) regions. Then, Ti/Pt is deposited immediately after the etching step through E-beam evaporation. The results, respectively after the resist development and metal lift-off are reported in Fig.5.3a and Fig.5.3b.



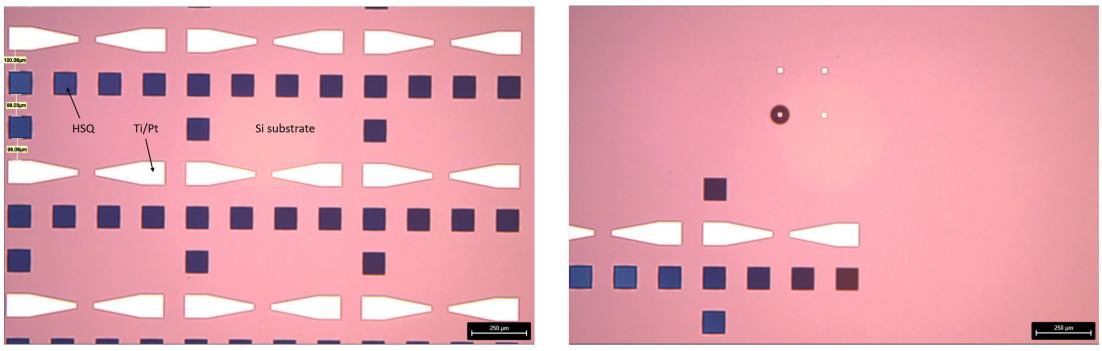
(a) **(b)**

Figure 5.3: Ohmic Ti/Pt contacts after development (a) and after lift-off (b).

The quality of the patterned features is excellent. This is also confirmed by the SEM picture of one of the Ti/Pt markers reported in Fig.3.32b (see Chapter 3.2.3).

5.2.2 HSQ Bond Pad Protection

HSQ is the negative resist used to prevent eventual damages caused by wire bonding to the thin gate oxide below the gate metal pads. Fig.5.4a illustrates the patterning of the HSQ protections perfectly aligned to the Ti/Pt contacts.



(a) **(b)**

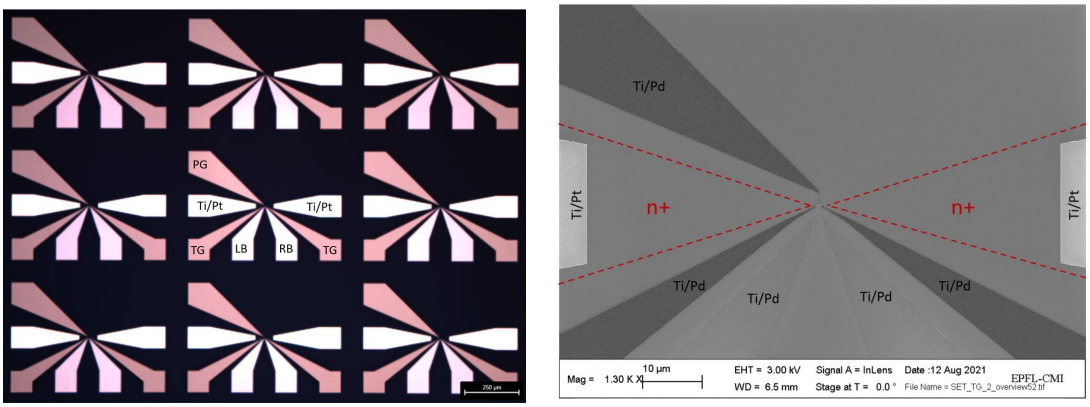
Figure 5.4: a) HSQ after exposure and developing. Being a negative resist, the exposed areas survive after development. b) HSQ survives also where the electron beam hits the resist looking for the marker.

5.2.3 Aluminum Oxide Blanket

A blanket oxide of 10 nm-thick Al_2O_3 is deposited over all the chip to prevent leakages from the long wires to the substrate. The technique employed is ALD at the temperature of 300°C. The resulting thickness, measured again by the ellipsometry, is equal to (10.42 ± 0.04) nm. Then, the Al_2O_3 is selectively etched away in a window of 10 μm × 10 μm in correspondence to the dot region. The etching tests confirm that a temperature of 60°C is necessary for wet etching of Al_2O_3 through H_3PO_4.

5.2.4 Palladium Gates and Interlayer Oxide

The two metal gate layers (barrier gates and top and plunger gate) are defined by EBL and lift-off and are separated by a 5 nm-thick interlayer oxide. The choice for this latter fell on Al_2O_3 because its higher dielectric constant (i.e, better insulation with respect to SiO_2 under the same thickness). However, the results shown at the end of this section suggest that SiO_2 may provide a better interface for the growth of the second metal layer. Fig.5.5a is an optical microscope image showing different SETs on the same chip with their routings after resist development. The zoom on the single device is reported in Fig.5.5b, taken with the SEM on the device after metallization.

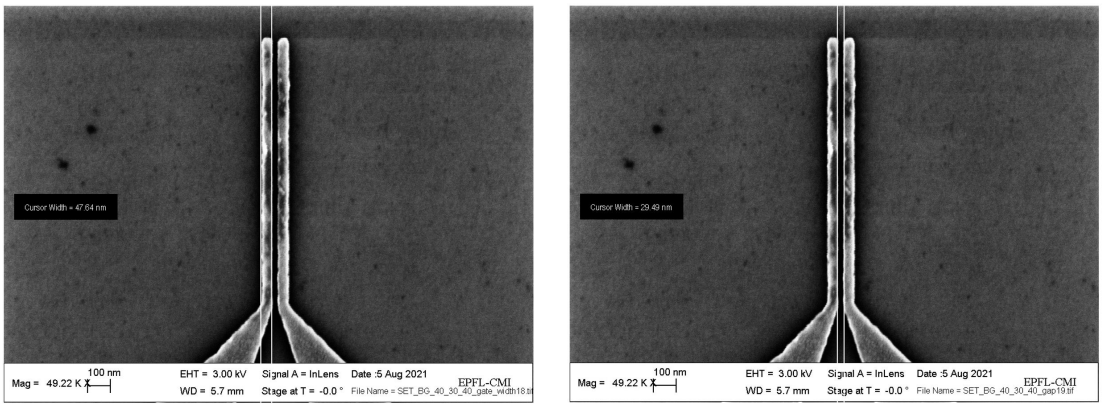


(a) **(b)**

Figure 5.5: a) Optical microscope image of the SET routings after development. Source and drain contacts are in Ti/Pt, TG is the top gate, PG the plunger gate and LG and RG the two barrier gates. b) SEM picture showing how the gates in the QD area are linked to the routings.

Barrier Gates

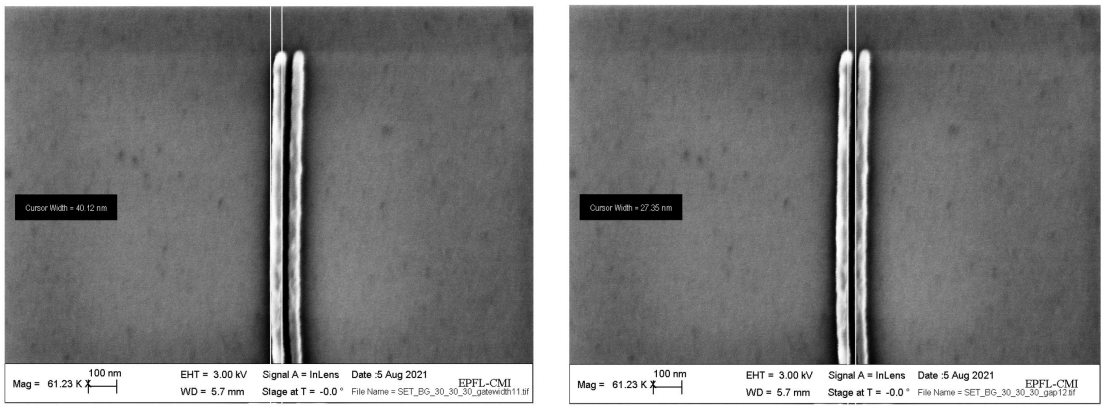
The first gate metal layer is constituted by the barrier gates. Fig.5.6 shows the results of 40 nm barrier gates separated by 30 nm of gap.



(a) **(b)**

Figure 5.6: SEM image of the 40-30-40 nm barrier gates. In a) the measured width is reported (47 nm), whereas in b) the measured gap (30 nm).

It can be noticed that the gates are slightly larger than expected (about 47 nm in this case), whereas the gap between them is the correct one (30 nm). The SETs with 30 nm gates separated by 30 nm of gap have been successfully patterned, with a resulting width of 40 nm and a gap of 27 nm, as shown in Fig.5.7.



(a) **(b)**

Figure 5.7: SEM image of the 30-30-30 nm barrier gates. In a) the measured width is reported (40 nm), whereas in b) the measured gap (27 nm).

This is at the moment the best obtained result in terms of resolution, corresponding of a dot size of ≈ 60 nm. In fact, the barrier gates of 30 nm separated by 20 nm of gap joined together (not reported here).

Interlayer Oxide

Al_2O_3 is used also as interlayer oxide, with a thickness of 5 nm. The deposition technique is again ALD at 300°C. The measured thickness by ellipsometer is equal to (5.41 ± 0.04) nm.

Top and Plunger Gate

The SET layout ends up with the top and plunger gates, reported in Fig.5.8. The

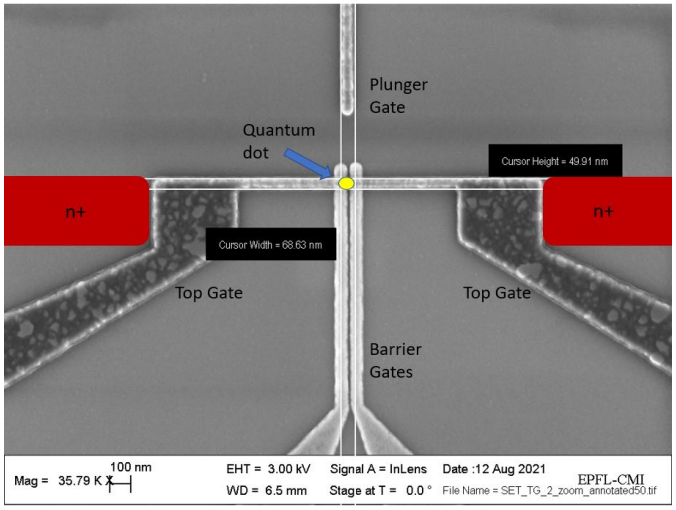
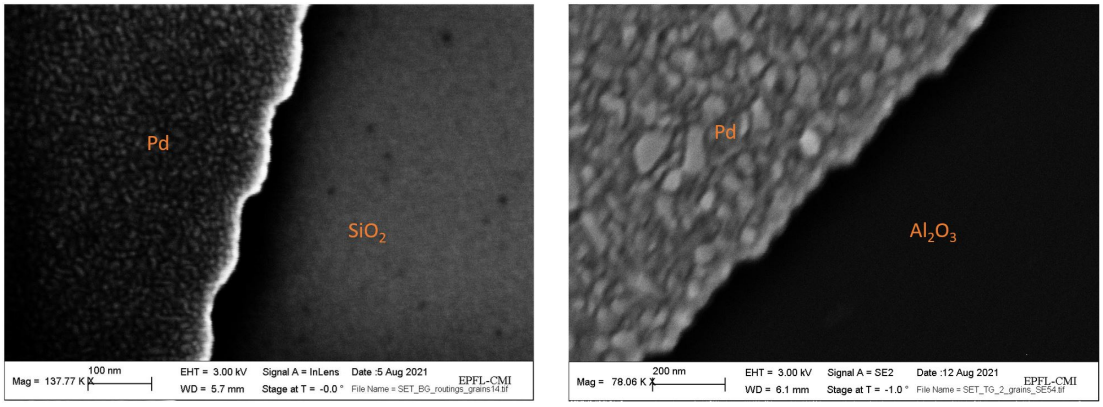


Figure 5.8: SEM image of the SET active region, showing the gate layers which define the QD. The long extension of the channel created by the top gate ($\approx 2\ \mu m$) is justified by the constraint in keeping the ohmic regions far from the dot.

top gate manifests some visible grains after the RTA at 400°C which, in principle, should improve the metal quality. This is probably due to the grains present in the interlayer Al_2O_3. In fact, the grains present in the amorphous oxide can be shaken by temperature, inducing a stress in the above layer. This problem cannot be solved by suppressing the annealing step, because this latter is necessary to reduce the trap density after the EBL step to pattern the top gate. The alternative to explore for the next fabricated batch is the ALD SiO_2. The origin of the grains must be independent on the deposition method, being it the same of the barrier gates, which in contrast exhibit an excellent uniformity. The comparison between the grains in Pd according to the underneath surface is reported in Fig.5.9.

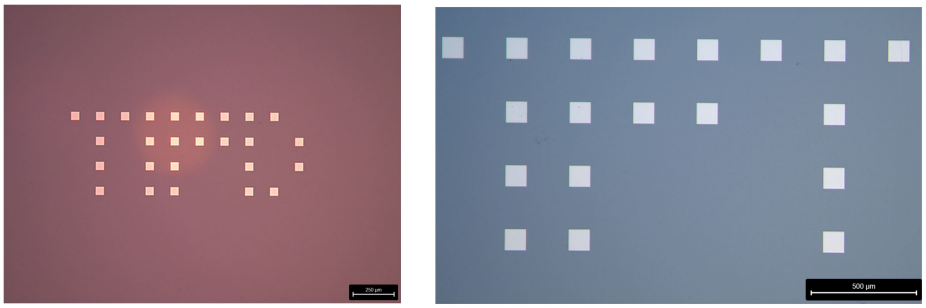


(a) **(b)**

Figure 5.9: SEM images of the Ti/Pd gates on SiO_2 (a) and on Al_2O_3 (b). Ti/Pd gates deposited on Al_2O_3 show larger grains.

5.3 MOS Capacitors

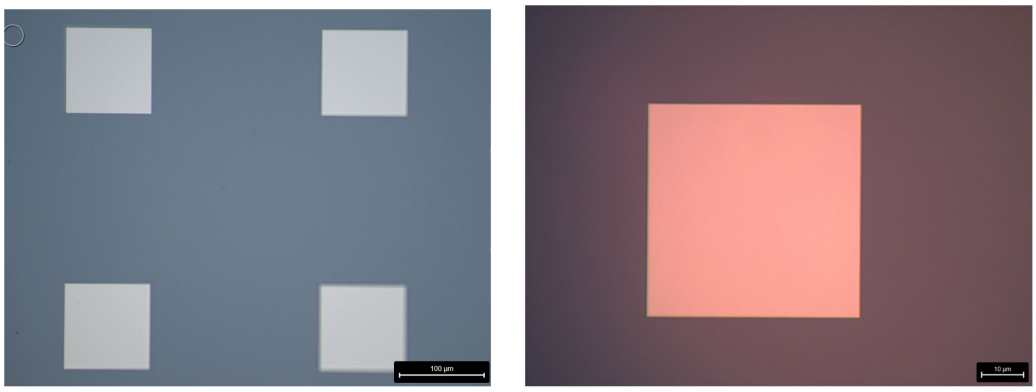
Several batches of MOS capacitors (MOScaps) have been fabricated for the study on the gate oxide (see Chapter 6.2). Regardless of the different oxidations and annealing steps performed, the MOScaps are fabricated through EBL (or PL), metal deposition (E-beam evaporation) and lift-off. Two different metals have been investigated for the top gate: Ti/Pd and Ti/Pt, i.e. the two used for the SET realization. A back gate was not realized. However, the oxide on the backside has been etched away. Two of the different MOScaps array are reported in Fig.5.10, showing both Ti/Pd (Fig.5.10b) and Ti/Pt (Fig.5.10b) MOScaps with an area of 100 μm × 100 μm.



(a) **(b)**

Figure 5.10: Ti/Pd (a) and Ti/Pt MOScaps (b) after lithography, metal deposition and lift-off.

Fig.5.11, instead, highlights the high quality of the lift-off for both the metals.



(a) **(b)**

Figure 5.11: Zoom on Ti/Pt (a) and Ti/Pd MOScaps (b) to show the quality of lift-off.

The oxide thickness in the different cases has been measured by the usual ellipsometer tool. Five different wafers have been oxidized for the study. The first two correspond to 5 and 10 nm-thick oxide grown with a standard dry oxidation at 1050°C. For the other three, the oxidation was carried out with DCE at 850°C. Then, a 5 s RTA at 1000°C has been performed on one of these last two 5 nm-thick oxide wafers, whereas for the other two oxides grown with DCE a 10 s RTA, always at 1000°C. The asked thickness of these last three wafers is of 10, 5 and again 5 nm. The resulting measurements are reported in Table 5.6.

5 nm standard (nm)	10 nm standard (nm)	5 nm DCE 5 s RTA (nm)	5 nm DCE 10 s RTA	10 nm DCE 10 s RTA
4.93 ± 0.04	9.73 ± 0.05	4.98 ± 0.04	4.61 ± 0.04	9.66 ± 0.05

Table 5.6: Ellipsometer measurements of the different oxidized wafers used for MOScaps fabrication.

5.4 Ohmic Contacts

The ohmic contacts are fabricated on chips from UTB FD SOI wafers already available in the Lab for other projects. These special SOI wafers have a top Si thickness of just 12 nm. The entire top Si film is doped with a donor concentration of 10^{19}. Being the top Si thickness compatible with SOI wafers used for the SET, they have been chosen to realize test structures for ohmic contacts. In fact, doping of the ohmic regions is not needed and these measurements can be performed before

shipping the SOI wafers for implantation, allowing to find the best parameters (see Chapter 4.2 and 6.3). The contacts, which consist in an array of pads of 100 μm $\times$ 100 μm, are first fabricated in Ti/Pd and then in Ti/Pt to greatly improve the quality. Standard photolithography, E-beam deposition and lift-off have been used to define the pads. Fig.5.12 shows an example of these pads arrays. Ellipsometer

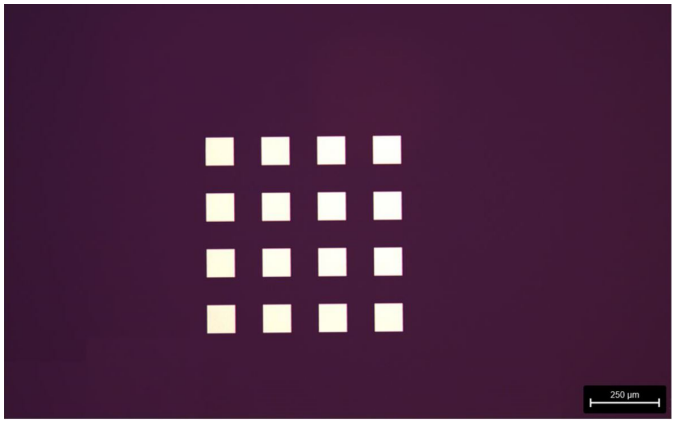


Figure 5.12: Ti/Pt contacts patterned on UTB FD SOI.

measurements reveal a top Si thickness of 11.05 nm, probably due to the formation of a native oxide.

Chapter 6

Gate Oxide and Ohmic Contacts Characterization

This chapter presents the electrical characterization results on the gate oxide and the ohmic Ti/Pt contacts on UTB FD SOI wafers. In the first case, C-V curves are registered for gate oxides grown in different conditions and before and after some annealing steps. The main goal is the extraction of trap density to find the best fabrication recipe for a ultra-high quality oxide. Traps extraction needs the knowledge of other parameters such as the flatband potential, oxide capacitance and thickness, semiconductor doping and workfunction and metal workfunction. All these parameters can be extracted from C-V curves measured for different oxide thicknesses and frequencies (see Appendix A). In the second case, measurements consist in I-V curves to test the quality of the ohmic contacts realized. From these measurements, it is possible to extract very important parameters like contact resistance, resistivity and doping (see Appendix B). Being the plasma doping adopted for the SOI SET ohmic regions a long process, the tests reported in this Thesis were not carried out on that wafers. Special UTB SOI chips with a top Si of 12 nm and uniform doping of 10^{19} cm^{-3}, already available at Nanolab, gave the possibility of doing preliminary tests. The aim is to find the best materials, fabrication processes and dopants concentrations for the future realization of the same contacts in the FD SOI wafers used for the SET.

6.1 Measurement Setup

All the measurements presented in this chapter are carried out with the aid of a probe station. In fact, to apply signals to the small metallic pads of semiconductor devices (100 $\mu m \times 100\ \mu m$ in the case of the tested devices), probes with a maximum radius of few decades of μm have to be used. A CASCADE probe station [68]

available in the Lab (see Fig.6.2), where both conventional and precision probes can be used for electrical measurements, is used for this purpose. Signals are generated and acquired by the Keithley 4200A-SCS parameter analyzer [69]. Connections between the probe station and the parameter analyzer are realized through SMU cables (*source measure unit*) for I-V curves, with both the *force* and *sense* terminals to ensure a precise measurement. On the other hand, for C-V curves additional CVU (*capacitance voltage unit*) boxes, to carry both the *low and high potential* and *low and high current*, have been used [70].

Figure 6.1: CASCADE probe station used for the electrical characterization. A microscope allows to check the good contact between probes and pads.

6.1.1 C-V Measurement technique

The technique employed by the Keithley parameter analyzer to measure capacitances consists in applying an AC small signal (about 25/30 mV) and measuring the resulting current. Then, the current is integrated over time to compute the charge and, from this latter, the capacitance can be easily computed for each DC voltage. In fact, the applied AC signal is superposed to a bias DC voltage, which is swept in the desired interval to obtain the C-V curve. This procedure is all integrated in the Keithley software. On the other hand, standard "LCR bridges" apply the same superposition of signals, but the value of capacitance is extracted from the measured impedance ($Z_C = 1/(j\omega C)$).
This measurement methods suggests that the presence of some series resistances

to the DUT (*device under test*) can alter the measurement by lowering the value of the measured capacitance. The first type of series resistances is constituted by the cables. Nevertheless, they can be easily compensated, directly by the Keithley software, through open and short compensations. The other series resistance is represented by the bulk of the semiconductor. Usually this is not a huge resistance, however it can start to be non negligible if an oxide is present on the backside of the wafer. In fact, when the wafer is oxidized, the oxide is grown also on the backside and, even if it was etched out in the MOS capacitors fabricated for the measurements reported in this chapter, a back contact was not realized through the immediate deposition of metal. As a consequence, the combination between the thin native oxide recreated on the backside with the absence of a metal back contact can greatly increase the series resistance. To overcome this problem, two expedients were adopted:

1. Silver paste was used on the backside of MOS capacitors to provide a better backside contact;

2. Analytical series resistances compensation was inserted in the Matlab script for parameters extractions according to the formulas present in Ref.[70]. These formulas allow to compute the compensated capacitance from the measured values of parallel conductance G_p, one for each DC voltage.

G_p and C_p (the MOS parallel capacitance) can be contemporary measured by both the electrometer through a phase-sensitive detection, which can discriminate between the in-phase component and the out-of-phase one [71]. In formulas:

$$v_o = \frac{R_L}{R_L + 1/(G_p + j\omega C_p)} v_i \approx (G_p + j\omega C_p) R_L v_i, \tag{6.1}$$

for small values of load resistance R_L, as it usually is. This means that the in-phase component is $v_o \approx G_p R_L v_i \sim G_p$ and the out-of-phase one is $v_o \approx \omega C_p R_L v_i \sim C_p$.

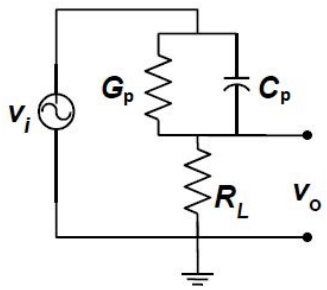


Figure 6.2: Circuit representation of the $C_p - G_p$ model for MOS capacitor [71]

6.2 Full study of the Gate Oxide

To study the gate oxide quality, quantified in the number of traps, MOS capacitors have been fabricated and tested. Different types of MOScaps have been realized for this purpose. The variable parameters are: the metal gate material (Ti/Pd and Ti/Pt), the oxide thickness (10 and 5 nm), dry oxidation with and without DCE, before and after a RTA at 400°C in forming gases and with and without a 1000°C RTA just after oxidation. Moreover, even if all the MOScaps are patterned through EBL, some of them have been fabricated by standard photolithography to understand the entity of the damages introduced by the high-energy electron bombardment during EBL.
From the measurement point of view, C-V curves measured with the Keithley parameter analyzer were taken at three different frequencies: 10 kHz, 100 kHz and 1 MHz to understand the f-dependent behaviour of interface traps. However, all the relevant parameters were extracted from the 1 MHz curves, being the less affected by distortion introduced by interface traps. In this way, the resulting $1/C^2$-V curves show a clearer shape and the flatband extraction from the depletion regime is more precise. For parameters extraction in each case study, all the C-V curves at the different frequencies were recorded for at least three MOScaps of the array. Moreover, all the curves are registered by introducing both a delay time of 100 ms and a hold time of 3 s to allow the MOScap to reach the equilibrium after each voltage step. In this way, the measurement is performed only when the capacitor is fully charged. These times have been chosen after the first experimental tests to reach a good trade-off between reliability and time necessary to get the measurement. The software Matlab has been used to plot all the curves and to extract all the parameters of this section.

6.2.1 First Results: Standard Dry Oxidation

The first set of MOScaps are fabricated on dry oxidized wafers according to the standard recipe used in CMi's cleanrooms for gate oxide growth. The temperature of the furnace is 1050°C. The results of the C-V curves for Ti/Pd MOScaps with both 5 and 10 nm-thick oxides are shown in Fig.6.3. By quick looking at the curves, it can be immediately noticed that traps contribute is non negligible. In fact, in an ideal MOS capacitor, i.e. with no traps of any type, the C-V curves should look the same at each frequency (larger than the minority carriers recombination and generation lifetime so that in inversion the capacitance does not rises to C_{ox}). However, it can be noticed that in proximity of the flatband voltage V_{FB} (which in a C-V curve can be qualitatively individuated as the point where the curve changes its concavity, being the potential linked to the second derivative of the

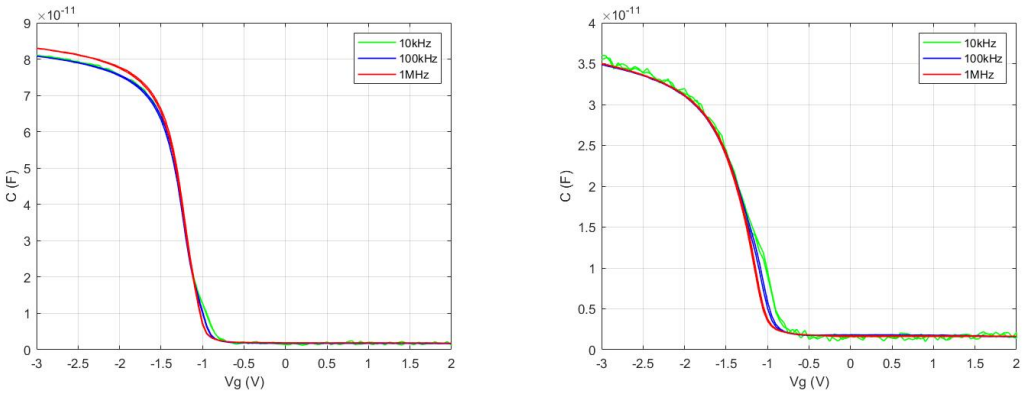


Figure 6.3: C-V curves of Ti/Pd MOS capacitors with a gate oxide thickness of 5 nm (a) and 10 nm (b). For these samples, the standard dry oxidation recipe was adopted.

charge density through the Poisson equation), the curves at lower frequencies show some "humps" because of the presence of interface traps. This effect is even more accentuated in the MOScaps with thicker oxide (10 nm), because the lower value of oxide capacitance forces the value of V_{FB} to be farther from the ideal one to compensate the presence of trap charges, according to Eq.(2.30). The parameters extraction passes through the $1/C^2$-V curves at 1 MHz, reported in Fig.6.4. $1/C^2$-V

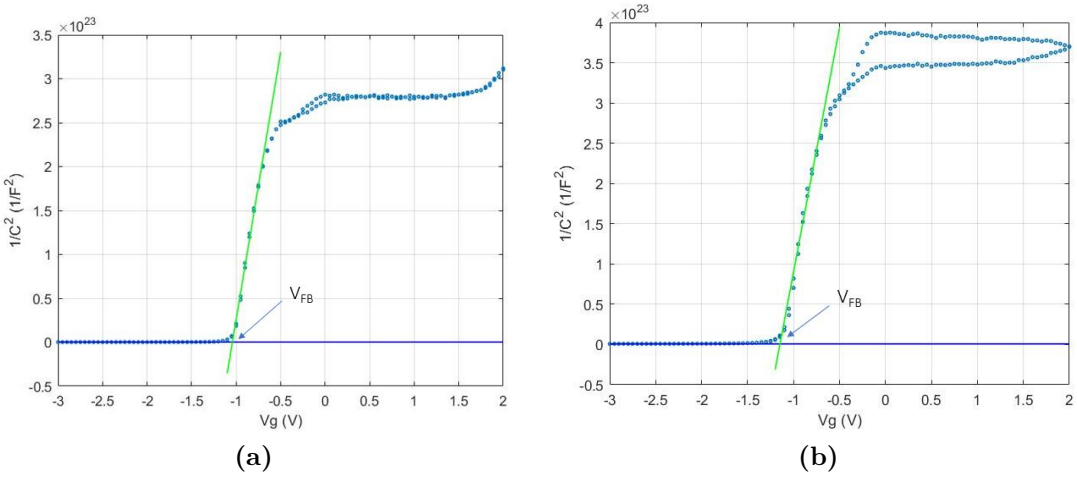


Figure 6.4: $1/C^2$-V curves of Ti/Pd MOS capacitors with a gate oxide thickness of 5 nm (a) and 10 nm (b).

curves highlight the presence of traps: V_{FB} is different for the two curves whereas, for an ideal MOScap, it should be imposed just by the workfunction difference (under the same substrate doping). Table 6.1 reports the extracted parameters for this first batch of MOScaps tested. V_{FB} and N_A are obtained through the extraction method reported in Appendix A, whereas ϕ_S from the combination of Eq.(B.2) and Eq.(2.2). Measures are carried out on three different devices, so only the average value is reported (standard deviations are negligible). Results of oxide

MOScap	5 nm	10 nm
C_{ox} (pF)	81.7	36.2
t_{ox} (nm)	4.22	9.55
$V_{FB}(V)$	-1.045	-1.202
N_A ($\times 10^{15}$ cm^{-3})	1.76	1.68
ϕ_S (V)	4.914	4.913

Table 6.1: Extracted parameters from both C-V and $1/C^2$-V curves. The samples are Ti/Pd MOScaps and the oxidation process is the standard dry oxidation for gate oxides.

thickness are consistent with the ellipsometry measurements reported in Chapter 5.3 and also the doping is in the expected range. In fact, this value corresponds to a resistivity of about 2.7 Ω·cm and the expected range is from 0.1 to 100 Ω·cm. The following step consists in extracting the trap density from the above obtained parameters. According to Eq.(2.30), the workfunction of the metal gate has to be perfectly known. However, this value is usually process dependent and, moreover, the metal used is a bilayer constituted by a Ti adhesion layer and Pd, so the exact value of ϕ_S cannot be known "a priori". Literature research [72] suggests that if the first metal layer is thick enough, usually more than few nm, the metal workfunction seen by electrons in the semiconductor is close to the one of the first layer. However, reported value for Ti workfunction differs of some fractions of eV (from 3.75 [72] to 4.33 eV [73]) and so, to get a precise value of traps concentration, an alternative approach has to be adopted. This approach consists in extracting simultaneously the workfunction difference ϕ_{MS} (which immediately gives ϕ_M if ϕ_S is known), taking advantage of the linear flatband voltage dependence on oxide thickness (see Appendix A). The results are shown in in Table 6.2 and in Fig.6.5. The extracted

ϕ_M	N_{eff}
3.99 eV	6.37×10^{11} cm^{-2}

Table 6.2: Extracted metal workfunction and total trap density.

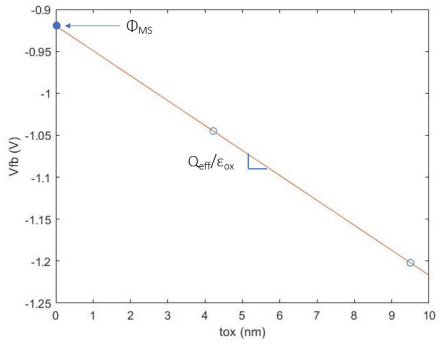
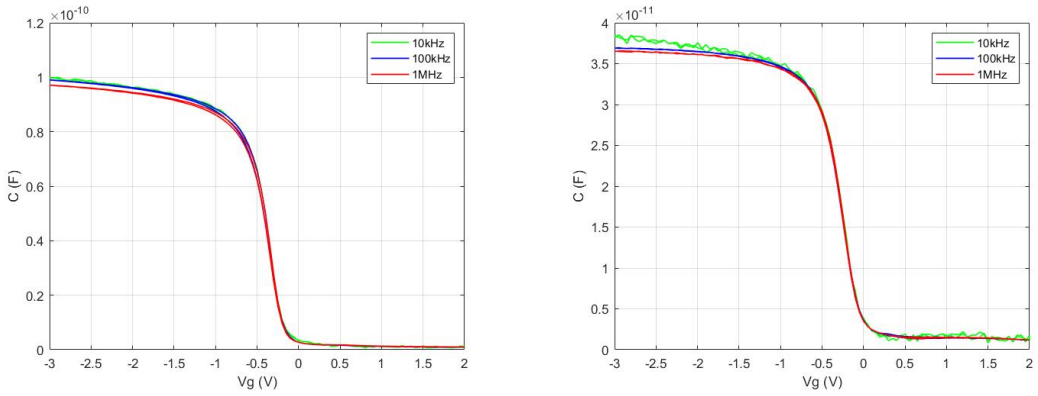


Figure 6.5: Flatband voltage as a function of oxide thickness. Intercepts gives the workfunction and the slope the traps charge (see Appendix A).

value of ϕ_{MS} is consistent with the results from literature [72],[73]. Notice that the trap density is the total contribute from the different types of traps. This value can be considered a good result in general, but not for a gate oxide, where this value should be in the order of 10^{10} cm^{-2}. The main contribute comes from interface traps, responsible for the "humps" of Fig.6.3. However, most of these traps can be passivated by a low temperature annealing in forming gases (FG) [22], [37].

Effects of Low Temperature RTA in FG

RTA in FG (5% H_2/95 % N_2) is the most effective mean to reduce interface traps. In fact, hydrogen can diffuse through the oxide and saturate the dangling bonds at the Si/SiO_2 interface. In this way, most of the interface states are effectively passivated. The best temperature for the annealing process is 400°C [37], whereas the time can vary from 10 to 30 minutes. In this work, 15 minutes are chosen, but, in principle, the longer the annealing time, the better is the passivation. Fig.6.6 illustrates the resulting C-V curves for the two stacks of MOScaps (5 and 10 nm) after the annealing step. The positive effects are visible just looking at the curves, which now look all the same at every frequency and for the two different oxide thicknesses. The "humps" disappeared and there is not any trace of hysteresis, neither in the 10 nm-thick MOScap. $1/C^2$-V curves (not reported here) allow again to extract all the relevant parameters, reported in Table 6.3. It can be immediately noticed that the V_{FB} values this time are very closed and separated by just some decades of mV, thus confirming an effective reduction in trap density. The V_{FB} vs t_{ox} curve reported in Fig.6.7 confirms this aspect and the extracted values of trap density and metal workfunction are reported in Table 6.4. The effective trap density is now almost in the desired order of magnitude, resulting in an acceptable result considering that EBL used to pattern the MOScaps creates traps because of electron bombardment, as well as the E-beam evaporator used to



(a) **(b)**

Figure 6.6: C-V curves of Ti/Pd MOS capacitors with a gate oxide thickness of 5 nm (a) and 10 nm (b) after the RTA in FG.

MOScap	5 nm	10 nm
C_{ox} (pF)	99.2	38.0
t_{ox} (nm)	3.48	9.10
$V_{FB}(V)$	-0.190	-0.153
N_A ($\times 10^{15}$ cm^{-3})	1.88	1.60
ϕ_S (V)	4.916	4.912

Table 6.3: Extracted parameters from both C-V and 1/C^2-V curves after RTA at 400°C in FG for 15 minutes.

deposit the Ti/Pd bilayer. This latter is responsible for X-rays generation which can damage the oxide (see Chapter 2.3.2). However, this result can be improved. Both a chlorinated oxidation and a high-temperature RTA just after oxidation are known to produce higher quality oxide and for this reason are explored in the next section. Nevertheless, before going on, few words are deserved for the other extracted parameters. First of all, the metal workfunction is quite different before and after the annealing step. The reason, confirmed by [6], is that before the RTA the workfunction is the one of the first metal layer (Ti) because the closest to the semiconductor. However, after the RTA step, some Ti atoms can diffuse into the Pd layer, thus modifying the effective workfunction seen by the electrons. The resulting $\phi_M = 4.70$ eV is a sort of average between the Ti and Pd workfunctions (5.22 eV). This result is extremely important because, for example, the threshold voltage of a MOSFET can depend on annealing steps if a bilayer gate is used. Moreover, it is also fundamental in engineering an ohmic contact,

ϕ_M	N_{eff}
4.70 eV	1.47×10^{11} cm^{-2}

Table 6.4: Extracted metal workfunction and total trap density after the RTA at 400°C in FG.

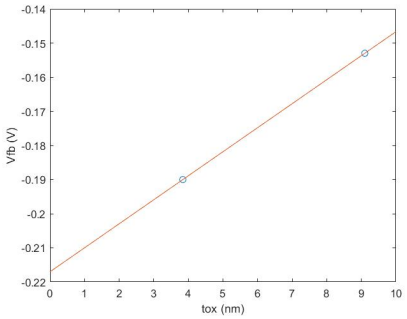


Figure 6.7: Flatband voltage as a function of oxide thickness after the RTA at 400°C in FG.

where the metal workfunction has to be known. This is the reason why also Ti/Pt MOScaps have been fabricated to compare the extracted workfunctions and prove why experimental results are better with Ti/Pt contacts. In conclusion, it can be noticed that the effective extracted oxide thickness is lower than before the RTA step. This may be due to two aspects:

1. Ti may diffuse also in the oxide, thus resulting in an effective dielectric thickness;

2. The hydrogen presence during the RTA step may force the diffusion of boron atoms present in the Si substrate into the oxide, which has a *segregation coefficient* m < 1. A < 1 segregation coefficient indicates the preference of the oxide in taking the impurities more than rejecting them from the interface and this effect is enhanced in H_2 atmosphere [9].

Even if the reason of the effective oxide thickness reduction is still not completely understood, it has to be taken into account when looking for the best oxide thickness. In fact, a 5 nm-thick gate oxide has been adopted with respect to a 10 nm-thick one to reduce charge noise, but electrical insulation has still to be guaranteed also after the annealing step. In the context of this work, a 5 nm gate oxide is sufficient considering that the oxide has this thickness only in the dot region (everywhere else there is also the thick Al_2O_3 blanket oxide) and that with a FD SOI substrate the voltages to apply to the barrier gates are very low.

6.2.2 Oxidation in DCE and high-T RTA after oxidation

To improve the obtained results, the oxidation is carried out in chlorine ambient. In this case, DCE is used to remove metal ions from the oxide and reduce interface traps [48]. The oxidation temperature is reduced to 850°C, because of the higher growth rate and the very thin oxide thickness requested. Just after oxidation, a 1000°C RTA of 5 s in FG was performed to reduce the fixed trapped charges [22],[37],[29],[32]. The resulting C-V curves are showed in Fig.6.8.

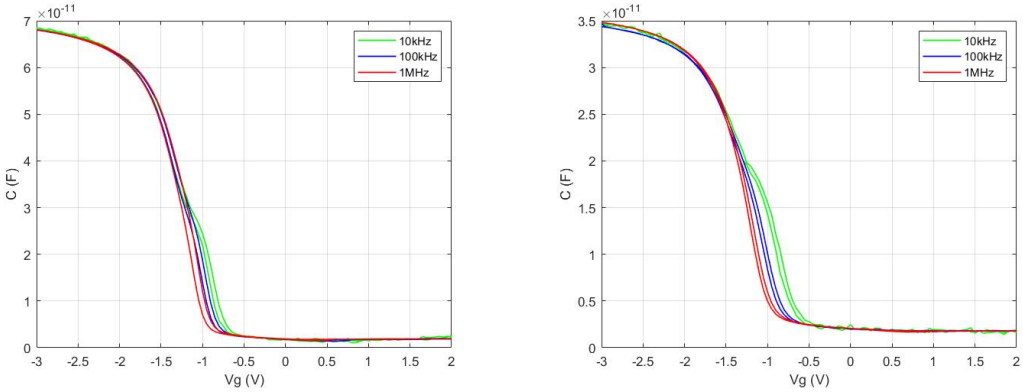


Figure 6.8: C-V curves of Ti/Pd MOS capacitors with a gate oxide thickness of 5 nm (a) and 10 nm (b). For these samples, dry oxidation with DCE, followed by 5 s RTA at 1000°C, was adopted.

MOScap	5 nm	10 nm
C_{ox} (pF)	68.1	34.4
t_{ox} (nm)	5.07	10.04
$V_{FB}(V)$	-1.016	-1.287
N_A ($\times 10^{15}$ cm^{-3})	2.63	3.42
ϕ_S (V)	4.925	4.932
ϕ_M (V)	4.19	
N_{eff} (cm^{-2})	1.18×10^{12}	

Table 6.5: Extracted parameters of MOScaps fabricated with oxidation in DCE and 5 s RTA at 1000°C.

The presence of "humps" and hysteresis in the C-V curves, as well as the largest difference in the flatband voltage between the 5 nm and 10 nm MOScaps remarks

a deterioration of the obtained results. The number of traps extracted through the usual method further confirms this aspect. The RTA at 400°C in FG helps reducing the trap density, even if the value is higher with respect to the one obtained with the standard oxidation. Fig.6.9 shows the resulting C-V curves, whereas Table 6.6 the extracted results. Since, on the contrary of what expected, this second

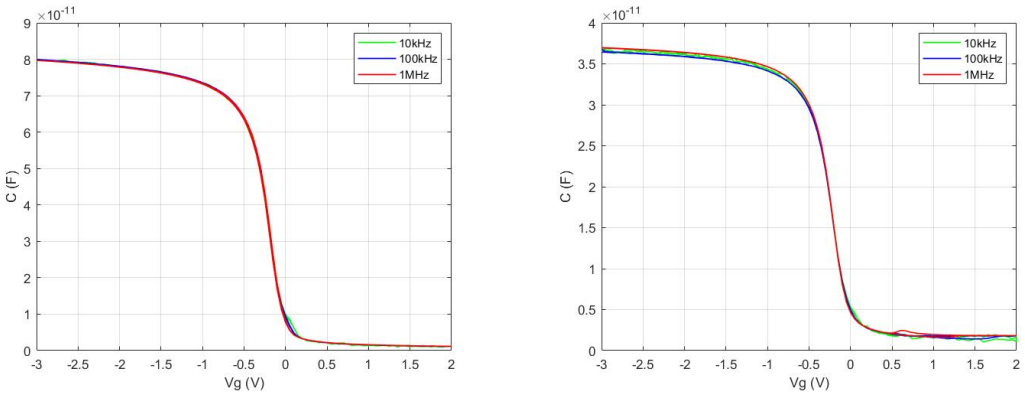


Figure 6.9: C-V curves of 5 nm (a) and 10 nm (b) MOScaps after the 15 min. FG RTA at 400°C.

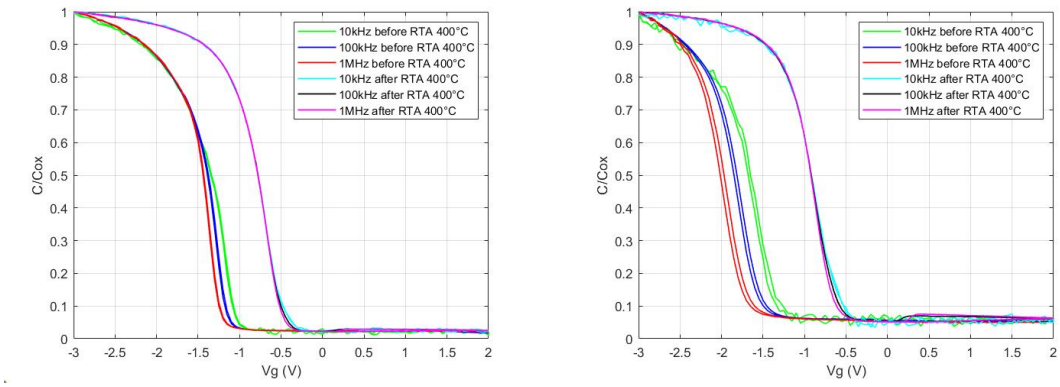
MOScap	5 nm	10 nm
C_{ox} (pF)	80.2	36.6
t_{ox} (nm)	4.30	9.42
$V_{FB}(V)$	-0.022	-0.120
N_A ($\times 10^{15}$ cm^{-3})	2.89	2.92
ϕ_S (V)	4.927	4.928
ϕ_M (V)	4.99	
N_{eff} (cm^{-2})	4.12×10^{11}	

Table 6.6: Extracted parameters after the 15 min. FG RTA at 400°C.

oxidation procedure generated more traps than the previous standard one, the physical origin to explain the results has to be found. However, in this situation, it is not possible to understand if they are due to DCE, to the high-temperature RTA or both. Strong suspicions fall on the high-T RTA, so new tests are carried out to confirm and quantify this aspect.

6.2.3 Standard Oxidation (no DCE) and high-T RTA after oxidation

MOScaps are now fabricated on the first batch of standard oxidized wafers (i.e., without DCE) after performing a 5 s RTA at 1000°C just after oxidation. In this way, it is possible to understand the responsible for worsening of the results presented in the last section. C-V curves of both the 5 nm thick and 10 nm thick gate oxide are plotted respectively in Fig.6.10a and in Fig.6.10b, showing both the results before and after the usual (and now approved) RTA at 400°C in FG for 15 minutes. Looking at the curves, one results is evident: V_{FB} is more negative than



(a) **(b)**

Figure 6.10: C-V curves of 5 nm (a) and 10 nm (b) MOScaps before and after the 15 minutes FG RTA at 400°C.

the previous case (with DCE). This observation indicates that the main responsible for the oxide degradation is the high-T RTA and not the DCE. The values reported

	before annealing		after annealing	
MOScap	5 nm	10 nm	5nm	10nm
C_{ox} (pF)	73.4	31.9	93.0	37.6
t_{ox} (nm)	4.70	10.8	3.71	9.17
$V_{FB}(V)$	-1.277	-1.814	-0.503	-0.704
N_A ($\times 10^{15}$ cm^{-3})	1.31	2.49	1.40	1.89
ϕ_S (V)	4.907	4.923	4.908	4.916
ϕ_M (V)	4.06		4.46	
N_{eff} (cm^{-2})	1.90×10^{12}		7.27×10^{11}	

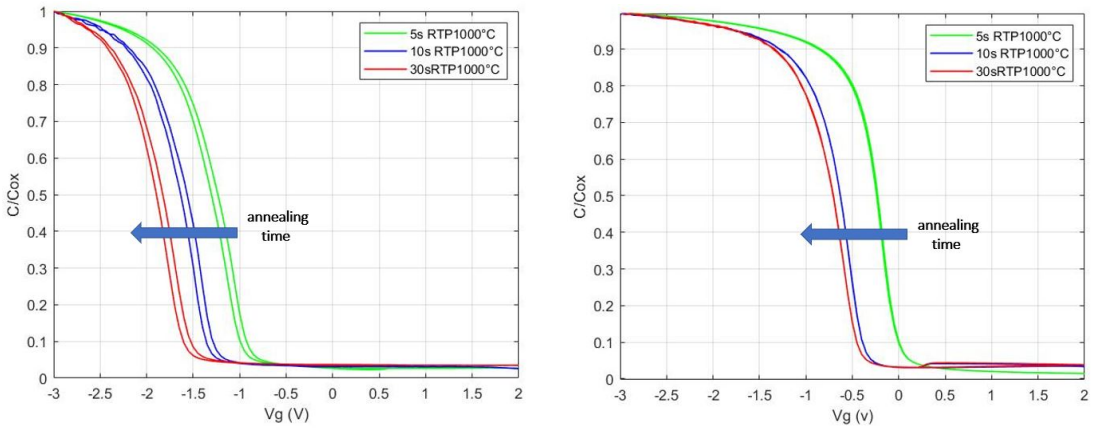
Table 6.7: Extracted parameters after the 15 minutes FG RTA at 400°C.

in Table 6.7 confirm the hypothesis made. In fact, the trap density extracted is the

worst reported value so far. This confirms that the high-T annealing is the main responsible for traps creation, whereas DCE removes traps. The reason why the high-T RTA is detrimental is not completely understood from the physical point of view. Nevertheless, two hypotheses are consistent with the results and find a response in the literature. First of all, the really high temperature can weaker the oxide, thus making it more sensitive to radiation such as E-beam litho used for the fabrication of the MOScaps [37]. In this way, the high-T annealing does not directly create traps, but makes EBL creating more traps than expected. Moreover, even if the annealing is performed in FG, 95% of the gas composition is constituted by nitrogen, which can diffuse into the oxide a depassivate the dangling bonds at the surface, thus creating more interface traps (see Section 2.3.2).

Effects of High Temperature RTA

To better study the detrimental effects of the high temperature annealing, other two batches of MOScaps have been fabricated. For these samples, the oxide was annealed, after oxidation in DCE, for respectively 10 and 30 s at 1000°C in FG. The reason of this study is motivated by the willing of further understand why the high temperature annealing, known for its property of reducing fixed trapped charges in the oxide, in the end created more traps than before. Moreover, since plasma doping is adopted for the doping of the ohmic regions, dopants need to be activated and a RTA at very high temperature is usually performed from 5 to 30 s. So, it is extremely important to understand what happens to the oxide during this step, in order not to degrade its quality or, at least, limit damages as much as possible. Only MOScaps with 5 nm oxide have been tested and the resulting C-V



(a) **(b)**

Figure 6.11: C-V curves before (a) and after (b) the 400°C RTA in FG.

curves before and after the usual 400°C RTA in FG step are reported in Fig.6.11. The reported curves are all recorded at 1 MHz. The green curves, i.e. the one corresponding to the 5 s RTA, comes from the capacitors described in Section 6.2.2. It can be noticed that there is a clear shift toward negative values of V_{FB} as the annealing times increases. This indicates that the longer the annealing, the more the traps created. The presence of hysteresis, combined to the fact that after the annealing at 400°C the blue and red curves are very close, confirm that the high-T RTA contributes to the creation of also interface traps, not only oxide trapped charges. The extracted values are reported in Table 6.9. These results lead to an

	before RTA 400°			after RTA 400°		
RTA 1000°C time	5 s	10 s	30 s	5 s	10 s	30 s
N_{eff} (cm^{-2})	1.18×10^{12}	2.63×10^{12}	3.74×10^{12}	4.12×10^{11}	2.54×10^{12}	2.75×10^{12}

Table 6.8: Extracted effective trap density as a function of the high-T annealing time.

important conclusion: an high-T RTA after the oxidation is absolutely detrimental for the quality of the oxide, especially if performed in nitrogen atmosphere. In order to take advantage of the high-T annealing to reduce fixed traps in the oxide, as indicated by literature, probably the annealing should be performed in complete hydrogen atmosphere and directly in the furnace after oxidation. As a consequence, particular attention has to be payed to the high-T annealing to activate the dopants, which has to be shortest and at the lowest temperature as possible. Ideally, a laser annealing should be the definitive adopted solution. In fact, the high power carried by the laser is able to activate the dopants in a so short time ($\approx ms$) that heat transfer is faster then the capability of the silicon layer to dissipate energy. This technology is present in IBM Zurich, one of the partners of the NCCR Spin project, so a future collaboration may be established for spin qubits fabrication.

Effects of oxidation with DCE

By comparing the results of Section 6.2.2 and 6.2.3, it is possible to quantify the benefits of DCE. This is because the only difference between the two batches of MOScaps is the oxidation procedure, respectively without and with DCE. Fig.6.12 shows the C-V curves at 1 MHz of 5 nm-thick oxide MOScaps with and without DCE. MOScaps fabricated on wafers oxidized with DCE show a less negative V_{FB}. The difference is of 0.147 V, which corresponds to a value of $\approx 7 \times 10^{11}$ cm^{-2} traps removed by DCE. This value, obtained from the usual Eq.(2.30) and taking into account the two different ϕ_{MS}, can be also simply obtained by subtracting the traps extracted from Table.6.5 and Table.6.7. On the other hand, it can be noticed that the curve corresponding to the DCE oxidized MOScaps shows more hysteresis than the other one. This may due to the fact that oxidation in DCE is carried out

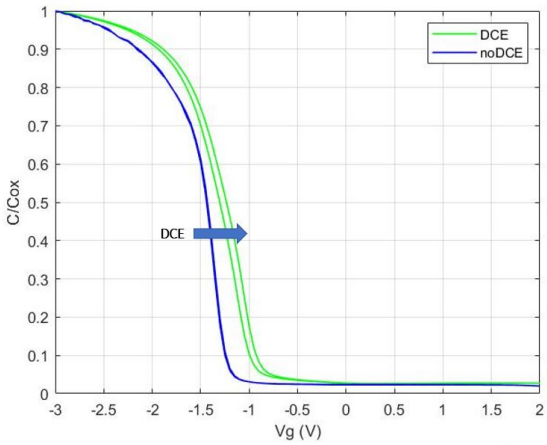


Figure 6.12: C-V curves of two different MOScaps without and with DCE.

at a lower temperature with respect to the standard dry oxidation without DCE (850°C vs 1050°C). As a consequence, probably more interface traps are present in the first case, but they are then mostly passivated by the low temperature RTA, as it is proven by the lower value of N_{eff} for the recipe with DCE also after the RTA at 400°C (4.12×10^{11} cm^{-2} vs 7.27×10^{11} cm^{-2}).

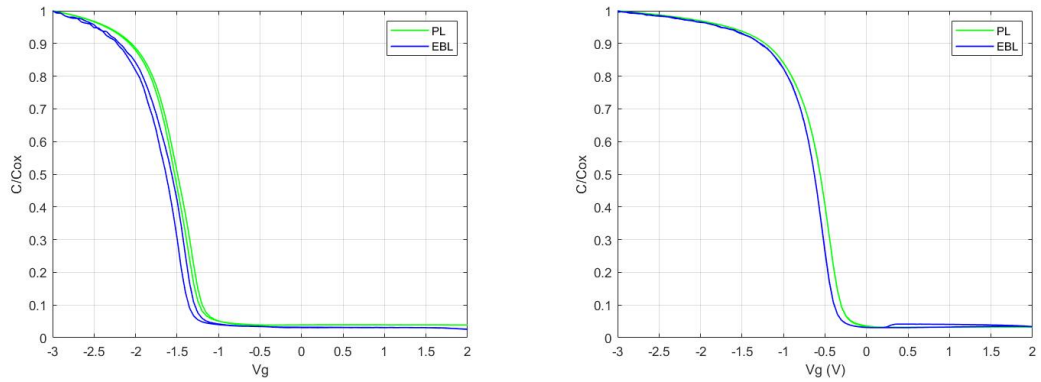
6.2.4 Best Fabrication Recipe

Under the light of the obtained results, the best fabrication recipe for a high-quality oxide consists in a dry thermal oxidation with DCE at 850°C, followed by a 400°C annealing in FG for 15 minutes (or, eventually, more), which can be also performed after every metallization step. To take advantage of the high temperature RTA ($>$ 800°C) to reduce fixed traps, the annealing should be performed in a complete hydrogen atmosphere and directly in the furnace just after oxidation. In fact, an annealing in nitrogen atmosphere or FG can increase the number of traps both in the oxide and at the interface. However, the annealing directly in the furnace has not been tested yet. In fact, it can be deduced from the obtained results that the best recipe obtained so far allows to reach the desired value of $\approx 10^{10}$ cm^{-2}. In fact, if the ideal value of flatband voltage is computed from the extracted values in the first tested MOScaps (standard oxidation): $V_{FB,ideal} = 3.99 - 4.91 = -0.92$ V, it can be noticed that this value is very close to the measured flatband voltage plus the contribute of DCE: $V_{FB,DCE} = -1.045 + 0.147 = -0.90$ V $\approx V_{FB,ideal}$. When the measured flatband voltage and the ideal one start to be closer than ≈ 0.02 V, for a 5 nm-thick gate oxide, traps are in the order of 10^{10} cm^{-2}. Moreover, this can be also proven comparing the trap density obtained with the standard recipe (6.37×10^{11} cm^{-2}) with the number of traps removed by DCE and extracted in the

previous section ($\approx 7 \times 10^{11}$ cm^{-2})), indicating that most of the traps present in these MOScaps can be removed by DCE, leading to the ideal value. Furthermore, this is confirmed also by comparing the trap density of Section 6.2.2 (oxidation with DCE and 5 s RTA at 1000°C), equal to 1.18×10^{12} cm^{-2}, to the value of traps introduced by the 5 s RTA at 1000°C, equal to (1.90 - 0.64)$\times 10^{12}$ cm^{-2} = 1.26$\times 10^{11}$ cm^{-2}. Also in this case, in fact, this value indicates that most of the traps present in these MOScaps are due to the 1000°C RTA and that, without this step, the ideal value could have been achieved. As a consequence, for the SOI wafers used for the SET fabrication, the oxidation was performed in DCE without any high-T annealing step.

6.2.5 Effects of E-beam Lithography

In order to understand the real damage caused to the oxide by EBL, on the same Si wafer were fabricated MOScaps both by EBL and standard photolithograhy. The wafer was oxidized in DCE and a 10 s RTA was performed to enhance the results, reported in Fig.6.13. The oxide thickness is 5 nm. For EBL-fabricated MOScaps,



(a) **(b)**

Figure 6.13: C-V curves of 5 nm-thick gate oxide MOScaps fabricated by EBL and standard PL before (a) and after (b) the RTA at 400°C.

	before RTA 400°		after RTA 400°	
litho technique	PL	EBL	PL	EBL
V_{FB} (V)	-1.293	-1.357	-0.378	-0.446

Table 6.9: Extracted effective trap density of MOScaps fabricated with the two lithographic techniques before and after the RTA at 400°C.

the flatband voltages are more negative and the C-V curve before the annealing shows more hysteresis than the ones fabricated by standard PL. This indicates, as expected, that EBL creates traps. The main goal is trying to understand how much worse the situation is after the RTA at 400°C with respect to PL. From the difference in V_{FB} and the oxide capacitance, it is possible to extract the difference of traps in the two cases (being the wafer and the metal the same, ϕ_{MS} is assumed the same for both batches of MOScaps). The computation gives $N_{eff} \approx 3 \times 10^{11}$ cm^{-2}, which is the effective number of traps introduced by EBL. However, if this result is compared with the extracted traps density of the MOScaps illustrated in Section 6.2.1 (no DCE, no high-T RTA), it can be noticed that the obtained value is higher in this last case, even if the MOScaps of that section are fabricated with EBL. This is because the MOScaps of this last study are fabricated on the wafer annealed at 1000°C for 10 s, thus confirming that the high-T RTA makes the oxide more sensitive to high-energy radiation. So, the difference in the number of traps between a MOScap fabricated by PL and EBL is expected to be minimum for devices fabricated on oxides not weakened by high-T RTAs. This is especially true after the 400°C RTA step, which, as shown in Fig.6.13b removes the hysteresis due to interface traps.

6.2.6 Ti/Pt MOScaps for Metal Workfunction Extraction

In the last sections, the workfunction of the Ti/Pd bi-metal layer has been extracted. After the RTA step, this value varies from 4.46 eV to 4.99 eV. The workfunction for heavily n+ doped Si is close to the electron affinity, i.e. 4.05 eV. When $\phi_M > \phi_S$ with a large barrier (in this case, from ≈ 0.40 to ≈ 0.95 eV), the metal-n-type semiconductor contact is a Schottky-type (see Appendix B). As a consequence, another metallization has to be explored for the realization of ohmic contacts on n+ doped Si. The choice falls on Ti/Pt. To extract the workfunction of such a bilayer, MOScaps with both 5 and 10 nm-thick oxide are fabricated. The starting wafer is oxidized with DCE, followed by a 10 s RTA at 1000°C. The resulting C-V curves are reported in Fig.6.14, whereas Table 6.10 shows the extracted results.

	before RTA 400°	after RTA 400°
ϕ_{MS} (V)	4.14	4.30
N_{eff} (cm^{-2})	3.56×10^{12} cm^{-2}	2.17×10^{12} cm^{-2}

Table 6.10: Extracted workfunction and effective trap density of Ti/Pt MOScaps before and after the RTA at 400°C.

The extracted workfunctions before and after the annealing are now closer to the semiconductor one, being however still larger. In theory, this situation would lead again to a Schottky barrier ($\phi_M > \phi_S$), but if the barrier height is low ($\approx 100 - 200$

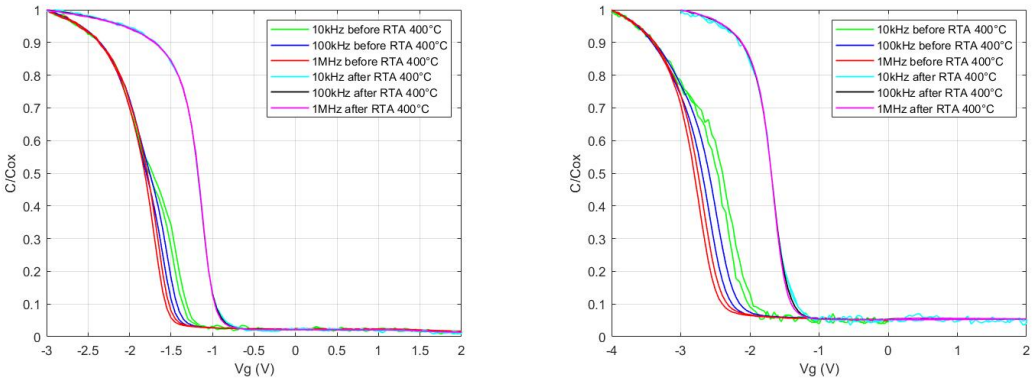


Figure 6.14: C-V curves of 5 nm-thick gate oxide MOScaps (a) and 10 nm-thick oxide (b) before and after the RTA at 400°C.

mV in this case), then the probability for an electron to tunnel through the barrier becomes not negligible anymore. Moreover, a lower barrier in energy makes also thinner the depletion region in the semiconductor, according to Eq.(C.3), resulting in a thinner tunneling barrier to cross by electrons. This also increases the tunneling probability, whose dependence on the barrier parameters is exponential. Moreover, when doping is high ($\geq 10^{19}$ cm^{-3}), current transport is definitively dominated by tunneling current and not by thermionic emission anymore, because of the depletion region width dependence on semiconductor doping (see again Appendix B). In this situation, the contact can be ohmic also in the case where $\phi_M > \phi_S$. Contacts characterization results reported in Section 6.3 demonstrate this last aspect. On the other hand, the extracted number of traps before and after the annealing is really consistent with the ones extracted for the same MOScaps, but fabricated with Ti/Pd (see Section 6.2.3).

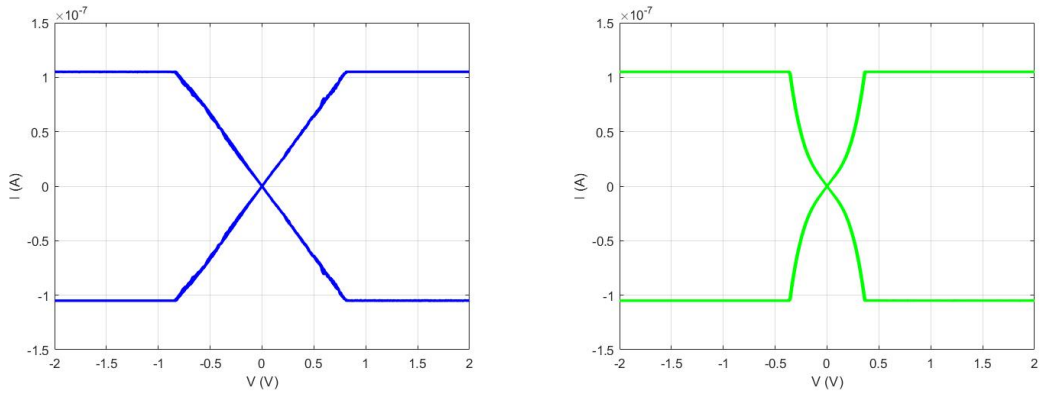
6.3 Ohmic Contacts on UTB FD SOI Characterization

The realization of ohmic, low resistance contacts on FD SOI wafer is one of the most critical steps of the process flow because of the very thin Si film (7 nm in one batch of SOI, 27 in the other one). First, this complicates the TCAD simulation for the doping of the ohmic regions (see Chapter 4.2) and then the proper choice of metals for the contact realization. The aim of this section is to provide information about the best parameters for the TCAD simulation (the target in the number

of donor atoms N_D to obtain with the simulation) and the metal to use. For this purpose, UTB SOI chips already present in the lab have been used as a substrate for the realization of the contacts and their characterization. The top Si thickness is of 12 nm, whereas all the chip is heavily doped, with a donor concentration of 1×10^{19} cm^{-3}. This allowed to carry out preliminary studies before shipping the SOI used for the SET realization to the external fab for plasma doping.

6.3.1 Ti/Pd Contacts

The first type of contacts is realized with Ti/Pd, being the metals already used for the gates. To characterize the contacts, simple I-V measures can be performed. Usually, to do these kind of measurements, test structures like the ones reported in Fig.4.5 (see Chapter 4.1.2) are used. However, in this case, being the whole substrate conductive, it is sufficient to apply a voltage difference between two nearby pads of the fabricated array. Then, by doing differential measurements, the channel resistance can be subtracted from the extracted value to get only the contact resistance. Also for these tests, measurements are carried out both before and after the RTA at 400°C for 15 minutes, in order to see eventual improvements or deterioration. Fig.6.15 shows the results of the I-V characteristic of two nearby



(a) **(b)**

Figure 6.15: I-V characteristic of Ti/Pd contacts before (a) and after (b) the RTA at 400°C.

pads before and after the annealing, where the maximum current has been set to 100 nA (the *compliance*). The n+ type Si is heavily doped, so its workfunction can be approximated to $\approx 4.05eV$. The Ti/Pd workfunction before the annealing extracted in the previous sections is around 4 eV. As a consequence, the contact is ohmic ($\phi_M < \phi_S$), whereas after the annealing, since the Ti/Pd workfunction increases to

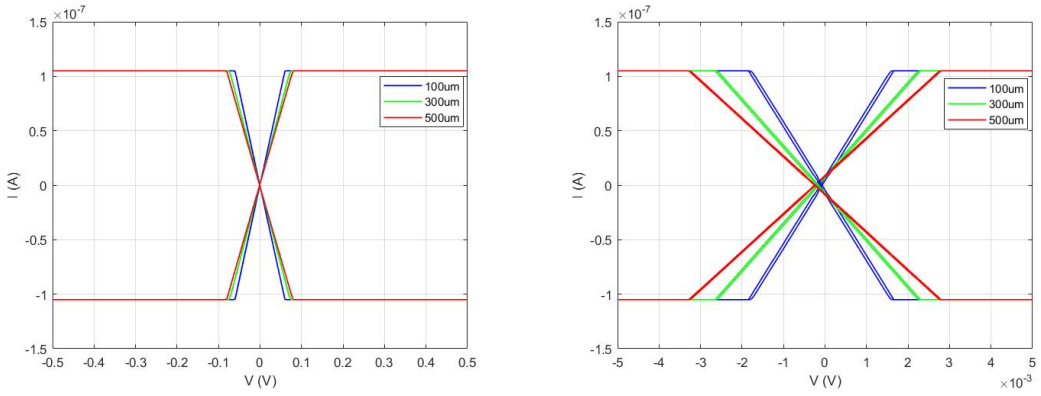
about 4.70 eV, the contact becomes of Schottky type. The results perfectly agree with the theory (see Appendix B). Nevertheless, two problems immediately come out. The first one is that, even if the contact is ohmic, the resistance is very high. The extracted value from the slope of the I-V curve is $R \approx 8$ MΩ. It is true that this value of resistance is not entirely due to the contacts, but also to the channel, which for a very thin Si layer can be very high. The channel resistance can be easily computed knowing the doping and the geometrical dimensions of the channel. In this case, the doping value of 1×10^{19} cm^{-3} corresponds to a resistivity $\rho = 5.9$ m$\Omega \cdot cm$. The channel is 100 μm long and wide and just 12 nm thick (see Chapter 5.4). With these values, the value of the channel resistance is equal to:

$$R_{ch} = \rho \frac{L}{W \cdot t} = 5.9 m\Omega \cdot cm \frac{100 \mu m}{100 \mu m \cdot 12 nm} = 4.9 k\Omega. \tag{6.2}$$

This highlights the fact that the huge resistance is due to the contacts and not to the channel. The second problem is the Schottky behaviour after the RTA. This is a huge problem because the RTA is a mandatory step to reduce the trap density. In conclusion, these results show that the use of Ti/Pd also for the contacts has to be abandoned toward other metals.

6.3.2 Ti/Pt Ohmic Contacts

The next metallization investigated consists in a Ti/Pt bilayer. The metal work-function values extracted in Section 6.2.6 predict better results, being closer to the silicon electron affinity. These results are confirmed by Fig.6.16. First of all, the



(a) **(b)**

Figure 6.16: I-V characteristic of Ti/Pt contacts before (a) and after (b) the RTA at 400°C.

first aspect to notice is that this time the contact is ohmic also after the RTA. Then,

focusing on the values of contact resistance, before the annealing it is still high. This aspect is proven by the fact that, increasing the channel length, the curves are still very close, indicating that the larger contribute to the total resistance is given by the contact resistance. The extracted value for the 100 um-long channel is of $R \approx 560\ k\Omega$, value which is very close also in the other two cases (300 and 500 um). Even though all the pads of the array are spaced by 100 um, if the current is measured between one pad and the second closest pads on the same line or column, the total channel length to be considered is of 300 um, not just 200 um. In fact, in this case, $R_C >> R_{ch}$ and so the electrons flowing from the first to the third pad do not pass through the metal of the second path, because the path offered by the channel is less resistive than the one offered by the metal-semiconductor contact crossed along the way. This is not completely true after the RTA, where the contact resistance starts finally to be comparable with the channel one. The three values of extracted resistance are 16.3, 23.3 and 28.8 kΩ for respectively the 100, 300 and 500 um-long channel. Looking at the values, it can be immediately noticed that the contact is still influencing the I-V characteristic. In fact, it the channel length is triplet, also the resistance should triple. However, this does not happen. Plotting the obtained resistance values as a function of the channel length allows to discriminate between the two components which give a contribute to the total resistance. In fact, if the points are interpolated with a line, the value of the intercept (corresponding to a channel length equal to zero) gives the contact resistance, whereas the slope the channel resistance per unit of length. Fig.6.17 reports the just mentioned curves. However, this time it is important to

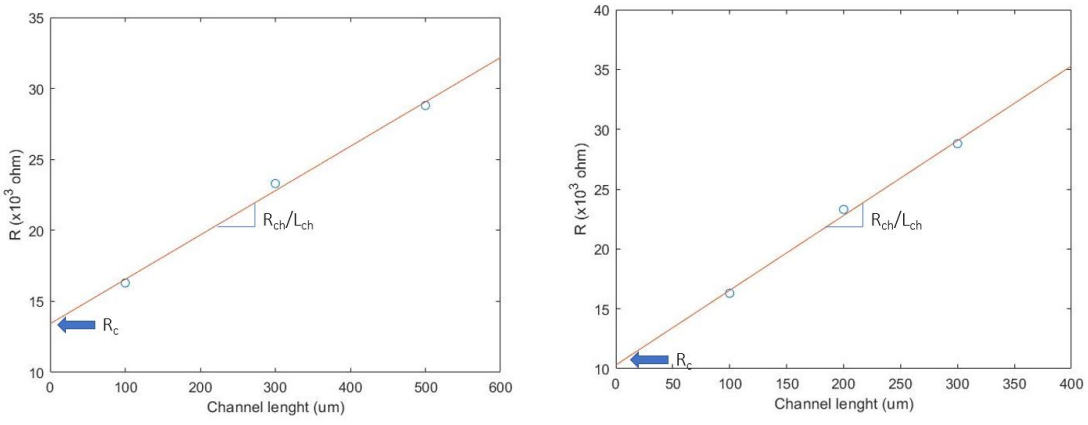


Figure 6.17: Total resistance as a function of the channel length. a) Limit for $R_C >> R_{ch}$. b) Limit for $R_C << R_{ch}$.

understand the real channel length to extract properly the parameters. Unluckily, the disposition of the fabricated contact pads does not help for this purpose. In fact, if the contact resistance starts to be comparable with the channel one, an electron flowing from the first pad to the third of the same line or row has two different conductive paths available. The first one is constituted by the 100 um long Si channel below the second pad, whereas the second by the one which starts with the contact between the semiconductor and the second pad, then continues in the metal (R $\approx$ 0) and ends again with the metal-semiconductor contact before going on to the third pad. This double pad can be modeled by two resistances in parallel, whose result is dominated by the lowest one between the two. To take into account this aspect, the R vs L_{ch} lines are plotted in the two limit cases. The one reported in Fig.6.17a represent the case in which $R_C >> R_{ch}$, meaning that the total channel length between the first pad and the third is equal to 300 um and 500 um between the first and the fourth. The second one, instead, represent the opposite situation, i.e. the one where $R_C << R_{ch}$. In fact, if the contact resistance is neglibile with respect to the channel one (as it should be in an ideal ohmic contact). the effective path that an electron has to follow inside the silicon is just of respectively 200 and 300 um between the first and the third pad and between the first and the fourth. Of course, in both cases there are no doubt concerning the shortest path between two nearby path, equal to 100 um. The extracted parameters lie in the window between these two extreme cases. Table 6.11 shows the extracted results in the two cases. The extracted value of doping confirms that the other parameters are

	R_c	R_{ch}/L_{ch}	ρ_{Si}	N_D
$R_C >> R_{ch}$	13.4 kΩ	31.3 Ω/um	3.75 $m\Omega \cdot cm$	1.7×10^{19} cm^{-3}
$R_C << R_{ch}$	10.3 kΩ	62.5 Ω/um	7.5 $m\Omega \cdot cm$	7.3×10^{18} cm^{-3}

Table 6.11: Extracted parameters of contact resistance, channel resistance per unit of length, resistivity and doping in the two limit cases.

correct. Unluckily, the value of R_C is still high. In fact, for an ideal ohmic contact this value should not exceed few fractions of Ω. However, in a SET the currents are usually in the order of hundreds of pA, whereas the applied source-drain voltages in the range of few mV. This implies that ohmic with a resistance of $\approx$ 10 kΩ experience a voltage drop of just few μV, which is negligible with respect to the applied voltages. However, since at cryogenic temperatures the contact resistance can be larger, to stay safe a doping of 1×10^{20} cm^{-3} has been chosen as a target for the TCAD simulations reported in Chapter 4.2, taking into account also their limitations in modeling doping in UTB FD SOI wafers.

Chapter 7

Conclusions and Outlook

The aim of this chapter is to summarize the most important results achieved during this Thesis and present the future prospective and plans.

7.1 Summary of the Obtained Results

Different goals have been achieved thanks to the work carried out during this Thesis:

- A well known process flow for the fabrication of QDs on a SiMOS platform has been extended to the SOI technology, whose benefits also in the quantum devices field have been mentioned here and may be demonstrated in the next years.

- Every step of the process flow is thought with always the same goal: charge noise minimization, from the oxide growth and annealing techniques to the choice of the materials for grains and non-uniformity reduction.

- The best oxidation technique for a high-quality and ultra-thin oxide has been individuated through C-V measurements on ad-hoc realized MOS capacitors. It consists in a dry thermal oxidation in DCE at 850°C. Sources of oxide degradation have been located in high temperature annealing steps (like the one for dopants activation) and E-beam lithography, whereas a low temperature annealing in FG for more than 15 minutes greatly reduces the number of interface traps at the Si/SiO_2 interface.

- Ti/Pt bi-metal layer has been found to be a good candidate for the realization of low resistance ohmic contacts on UTB FD SOI wafers after a proper annealing step.

- TCAD simulations for plasma implantation have been optimized to reach the desired dopants concentration in the thin top Si layer without damaging the BOX.
- Last, but not least, all the theory describing the quantum dot realization in semiconductor structures, the SET working principle, quantum dot bias spectroscopy, charge noise characterization and possible origin due to traps has been deeply investigated and make this work suitable for people interested in enlarging their knowledge in this emerging field.

7.2 Future Plans

The work presented in this Thesis represents only the very beginning of a huge and ambitious project whose goal is the realization of high-fidelity electron spin qubits on the SOI platform explored here. Clear advantages in charge noise reduction, decoherence and relaxation times, speed, scalability etc. with respect to the other competitors have to be demonstrated with the future work.
Under this light, the main future steps consist in:

- Concluding the SET fabrication on the implanted SOI wafers.
- Testing the SET at cryogenic temperature to demonstrate Coulomb blockade and extract fundamental parameters such as charging energy, dot capacitance, dimensions and lever arm.
- Characterizing the charge noise in the mK regime and comparing it with the actual state-of-the-art semiconductor QDs technologies.
- Fabricating the QDs array integrated with micromagnets to realize a physical platform for electron spin qubits.
- Characterize the electron spin qubits and compare the results.

The project is ambitious because of the strength of the contender platforms, mostly represented by hole spin qubits in Si/Ge quantum wells and nanowires, planar Ge and Si FinFETs. Every platform has its advantages and drawbacks, but the SOI SiMOS one presented in this Thesis is very versatile and can be adopted to hole spin qubits by changing the doping of the ohmic regions. In principle, also both electron and hole spin qubits can be realized on the same wafer and compared. In conclusion, independently on which one of these candidates will triumph over the others, the day when quantum computers will be finally useful for daily life is always closer and I am really honored and excited to be part of a large scientific community made by people who works very hard every day to realize this dream.

Appendix A

MOS Capacitor Theory for Oxide Characterization

This Appendix is dedicated to the mathematical equations which describe the behaviour of the MOS capacitor, in order to properly understand the experimental results reported in Chapter 6 and the extraction methods of the desired parameters from C-V curves.

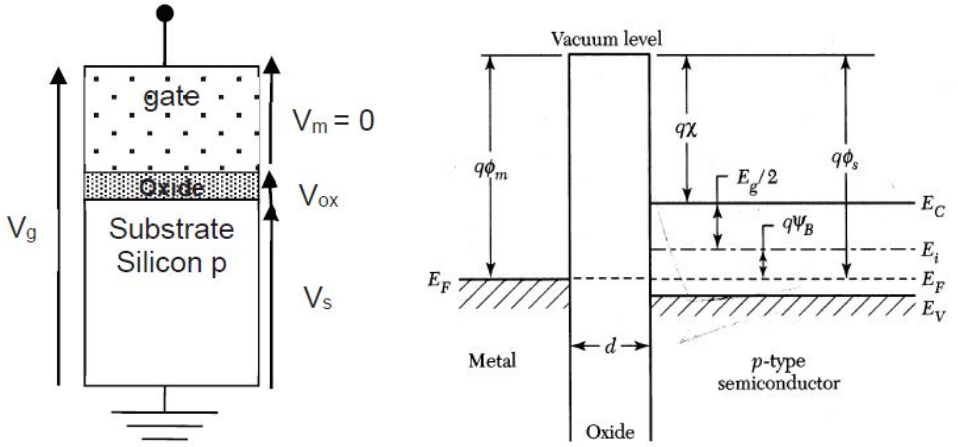
A.1 MOS Equations

In a MOS capacitor, a thin oxide layer, called gate oxide, separates a top metallic gate from the bulk semiconductor, as shown in Fig.A.1a [71]. If a gate voltage V_G is applied, the voltage drop across the structure, assuming the metal ideal, is so distributed:

$$V_G - V_{FB} = V_{ox} + V_s, \tag{A.1}$$

where V_{ox} is the voltage drop across the oxide, V_s the one across the semiconductor (the surface potential ψ_S in the main text) and the flatband voltage is the difference between the metal workfunction and the semiconductor one. In all the experiments carried out in Chapter 6, the MOS capacitors are fabricated on a p-type substrate. In this case, the Si workfunction is the sum of the electron affinity χ_e, half of bandgap E_g and the bulk potential ψ_B given by Eq.(2.2). ψ_B is the distance between the intrinsic Fermi level E_{FI} and the Fermi level E_F, which in a p-type semiconductor lies between E_{FI} and the valence band E_V, as shown in Fig.A.1b [22] and through the equation:

$$q\phi_S = \chi_e + \frac{E_g}{2} + \psi_B. \tag{A.2}$$



(a) **(b)**

Figure A.1: a) MOS capacitor schematic. Assuming the metal ideal (no voltage drop across it) the gate voltage applied is all distributed between the oxide and the semiconductor [71]. b) p-type MOS capacitor band diagram before contact [22].

The voltage drop across the oxide can be written as:

$$V_{ox} = \frac{Q_{tot}}{C_{ox}}, \tag{A.3}$$

where Q_{tot} is the total charge at the oxide-semiconductor interface (per unit of area) and C_{ox} is the oxide capacitance (again, per unit of area), expressed as:

$$C_{ox} = \frac{\varepsilon_{ox}}{t_{ox}}, \tag{A.4}$$

being t_{ox} the oxide thickness. C_{ox} is a fixed parameter depending only on the technology, not on operating conditions. However, the total capacitance of the MOS system does, being the series of two capacitances: the oxide capacitance C_{ox} and the semiconductor capacitance C_s, this latter being voltage dependent. The following computations demonstrate this last aspect (as a remind, for series capacitances, it is summed up the reciprocal of the different contributes) :

$$C_{tot} = \frac{dQ_{tot}}{dV_G} = (\frac{dV_G}{dQ_{tot}})^{-1} = (\frac{dV_{ox}}{dQ_{tot}} + \frac{dV_s}{dQ_{tot}})^{-1} = (\frac{1}{C_{ox}} - \frac{dV_s}{dQ_s})^{-1} = (\frac{1}{C_{ox}} + \frac{1}{C_s})^{-1}, \tag{A.5}$$

where $\frac{dV_{FB}}{dQ_{tot}} = 0$, $C_s = -\frac{dQ_s}{dV_s}$ and Q_s is the semiconductor charge, equal for definition to $-Q_{tot}$ and constituted by two different contributes: the depletion charge Q_d and the inversion charge Q_n, according to the operating regime of the device. In general:

$$Q_s = Q_d + Q_n. \tag{A.6}$$

In conclusion, in a MOS capacitor the C-V curve shows a voltage dependence because of the voltage dependence of the semiconductor capacitance. To recap:

$$\frac{1}{C_{tot}(V_G)} = \frac{1}{C_{ox}} + \frac{1}{C_s(V_G)}. \tag{A.7}$$

To understand the shape of the C-V curves reported in Chapter 6, the capacitance in the different operating regimes has to be analyzed.

Accumulation ($V_G < V_{FB}$):

In accumulation, holes are accumulated at the interface (always considering a p-type MOS capacitor), so there are no depletion charges nor inversion ones. As a consequence, the value of the total capacitance is close to the oxide capacitance:

$$C_{tot,acc} = C_{ox} = \frac{\varepsilon_{ox} A}{t_{ox}}, \tag{A.8}$$

where A is the total area of the MOS capacitor.

Depletion ($V_{FB} < V_G < V_{th}$):

In depletion, the majority carriers (i.e., the holes) are depleted from the interface as a consequence of the always more positive gate voltage applied. As holes are depleted, in the depletion region, called also space charge region, a net negative charge results from the negatively charged acceptor dopants. The depletion region becomes an insulating region for all intents and the result is the same of having a thicker oxide. As a consequence, the total capacitance drops from C_{ox} to its minimum value, the one for which the depletion region is maximum. To find the minimum value of the depletion capacitance, first an analytical expression for the width of the depletion region, denoted as x_d, has to be found. The charge density per unit of volume ρ in the depleted region is equal to $\rho_d = -qN_A$. Integrating twice the charge density, through the Poisson equation, it is possible to extract the width of the depleted region x_d as a function of the voltage drop across the semiconductor V_s:

$$x_d = \sqrt{\frac{2\varepsilon_{Si} V_s}{qN_A}}. \tag{A.9}$$

This expression of x_d allows to compute the depletion capacitance as a function of V_s:

$$C_d = \frac{\varepsilon_{Si}}{x_d} = \sqrt{\frac{q\varepsilon_{Si} N_A}{2V_s}}. \tag{A.10}$$

This equation underlines the fact the, by increasing the gate voltage (and so V_s), the depletion capacitance drops and, as a consequence, also the total capacitance

C_{tot} does. Since the width of x_d is proportional to $\sqrt{V_s}$, the maximum value of x_d occurs for the maximum voltage drop across the semiconductor, i.e. $2\psi_B$. In fact, for this value, the population of majority carriers, which at the interface ($x = 0$) and for $V_s = 0$ is equal to N_A, is completely inverted with the minority carriers population, thus becoming equal to $\frac{n_i^2}{N_A}$, according to the equations [71]:

$$p(x = 0) = N_A exp(-\frac{qV_s}{k_B T}), \tag{A.11}$$

$$n(x = 0) = \frac{n_i^2}{N_A} exp(\frac{qV_s}{k_B T}). \tag{A.12}$$

As a consequence, for this value of V_s, the width of the depletion region does not increase anymore and any further applied voltage drops all across the oxide. The maximum depletion region is thus equal to:

$$x_{d,max} = \sqrt{\frac{2\varepsilon_{Si}(2\psi_B)}{qN_A}}. \tag{A.13}$$

And the corresponding minimum capacitance becomes:

$$C_{min} = C_{tot|x_{d,max}} = (\frac{1}{C_{ox}} + \frac{1}{C_{d,min}})^{-1} = (\frac{t_{ox}}{\varepsilon_{ox}} + \frac{x_{d,max}}{\varepsilon_{Si}})^{-1} = \frac{\varepsilon_{ox}}{t_{ox} + \frac{\varepsilon_{ox}}{\varepsilon_{Si}} x_{d,max}}. \tag{A.14}$$

The general expression of the depletion capacitance as a function of V_G, is now demonstrated. This aspect will be very useful to understand how C-V curves can be used to extract parameters like the flatband voltage and the substrate doping. In the depletion region, the only net charge is given by the depletion charge (per unit of area): $Q_d = -qN_A x_d = -Q_{tot}$. As a consequence, the voltage drop across the oxide can be written as $V_{ox} = \frac{qN_A x_d}{C_{ox}}$. Now combining this last equation with Eq.(B.1) and Eq.(B.9), the result is:

$$V_G - V_{FB} = \frac{qN_A x_d}{C_{ox}} + \frac{qN_A x_d^2}{2\varepsilon_{Si}}. \tag{A.15}$$

This is a second order equation in x_d, whose solution (the acceptable one) is:

$$x_d = -\frac{\varepsilon_{Si}}{C_{ox}} + \sqrt{(\frac{\varepsilon_{Si}}{C_{ox}})^2 + \frac{2\varepsilon_{Si}}{qN_A}(V_G - V_{FB})}. \tag{A.16}$$

Substituting Eq.(B.16) into Eq.(B.10) gives the depletion capacitance as a function of the gate voltage. resulting in a total capacitance equal to:

$$C_{tot,depl} = \frac{C_{ox}}{\sqrt{1 + \frac{2C_{ox}^2(V_G - V_{FB})}{q\varepsilon_{Si}N_A}}}. \tag{A.17}$$

Inversion ($V_G > V_{th}$):

When V_s equals ψ_B, the concentration of both majority and minority carriers is the same and equal to n_i, always according to Eq.(B.11) and (B.12). For higher values of V_s, inversion between minority and majority carriers takes place. For $\psi_B < V_s < 2\psi_B$, the concentration of minority carriers is still negligible and for this reason this regime is called *weak inversion*. In this condition, the depletion of holes is still occurring from n_i to the minimum value of $\frac{n_i^2}{N_A}$, reached for $V_s = 2\psi_B$. As a consequence, the predominant capacitance in this regime is still the depletion one. However, when V_s reaches $2\psi_B$, i.e. when V_G is equal to the threshold voltage given by Eq.(2.1), the width of the depleted region cannot increase anymore and the total capacitance remains fixed to C_{min}. However, this is not always true. In fact, for this to be true, the inversion layer has to be confined at the edge of the depleted region. This is what happens when the frequency is sufficiently high, but if the frequency is lower than the generation-recombination rate of minority carriers at the surface of the depleted region, they are able to follow the AC signal, leading to charge exchange with the inversion layer in step with the measurement signal [22]. As a consequence, the effect of the depleted capacitance is bypassed and the capacitance comes back to C_{ox}. Usually this happens for f<100 Hz, a condition which is never fulfilled in all the measurements carried out during the Thesis and, for this reason, all the C-V curves reported in Chapter 6, have a shape similar to the one reported in Fig.A.2a [70].

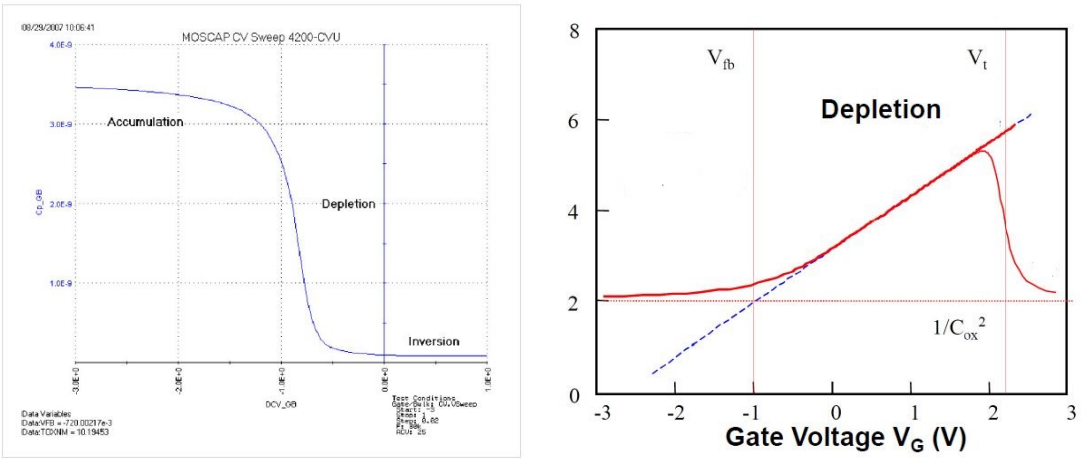


Figure A.2: a) C-V curve at high frequency highlighting the three operation regimes [70]. b) $1/C^2$-V curve. In depletion, the curve shows a linear dependence useful for V_{FB} and doping extraction [71].

A.2 Parameters extraction from C-V curves

C-V measurements allow to extract several useful parameters such as oxide capacitance, oxide thickness, flatband voltage, threshold voltage, doping concentration, metal workfunction, oxide and interface traps and others. C_{ox} can be extracted from the value of capacitance in strong accumulation, i.e. where it is constant and equal to C_{ox}. Then, the oxide thickness can be extracted from this value through:$t_{ox} = \frac{\varepsilon_{ox}}{C_{ox}}$.
Then, the extraction of the doping concentration and the flatband voltage can be performed in different ways. The one proposed in this Thesis is the $1/C^2$ method. This method consists in plotting the $1/C^2$ vs V_G curve, simply from C-V measured values. In fact, taking the square of Eq.(B.17), it can be noticed that the $1/C^2$ vs V_G curve is linear in depletion, as shown also in Fig.A.2b:

$$\frac{1}{C_{tot}^2} = \frac{1}{C_{ox}^2} + \frac{2}{qN_A\varepsilon_{Si}}(V_G - V_{FB}). \tag{A.18}$$

Eq.(B.20) is the equation of a line, whose intercepts with $1/C_{ox}^2$ gives the value of V_{FB}, whereas the slope gives the value of the doping concentration:

$$N_A = \frac{2}{q\varepsilon_{Si}} \frac{1}{\frac{\Delta(1/C_{tot}^2)}{\Delta V_G}}, \tag{A.19}$$

being C_{tot} the total capacitance per unit of area. Once the doping concentration is known, the semiconductor workfunction ϕ_S can be computed through Eq.(B.2). Then, if the trap concentration was negligible, from the measurement of the flatband voltage, the metal workfunction ϕ_M would be directly extracted. In a similar way, if the metal workfunction was perfectly known, the distance between the ideal value of V_{FB}, given by $\phi_M - \phi_S$, and the extracted one, would give directly the trap charge, according to Eq.(2.30). Then, from the trap charge, the traps concentration can be computed directly by: $N_{eff} = Q_{eff}/q = (N_{it} + N_f + N_{ot} + N_m)/q$. However, this is rarely the case. In fact, the metal workfunction is in most of the cases process-dependent and so a contemporary extraction of both ϕ_{MS} and Q_{eff} is needed to ensure correct results. The solution adopted in this thesis to overcome this problem consists in taking C-V curves for different oxide thicknesses [74]. In fact, if this is done, Eq.(2.30) can be rewritten as:

$$V_{FB} = \phi_{MS} + \frac{Q_{eff}}{\varepsilon_{ox}} t_{ox}, \tag{A.20}$$

which is the equation of a line in t_{ox}. As a consequence, V_{FB} values taken at different t_{ox} allow to draw a line whose intercepts gives directly ϕ_{MS} and the slope Q_{eff}.

Appendix B

The Metal-Semiconductor Contact

The contact between a metal and a semiconductor is one of the most diffused example of electrical junctions. It can be of two types: rectifying (called also *Schottky*) or linear (called also *ohmic*).

B.1 Schottky Barriers

Considering a n-type semiconductor, a Schottky barrier is created when the metal workfunction ϕ_M is larger than the semiconductor one ϕ_S. As a consequence, to establish the thermal equilibrium, when the two materials are in contact, electrons flow from the semiconductor to the metal, creating a depletion region. The unbalanced dopants result in a positive net charge, which originates an electric field and a potential barrier. This barrier prevents further electrons to flow from the semiconductor to the metal, like in a p-n junction at thermal equilibrium. This equilibrium situation is depicted in Fig.B.1.a [22]. Exactly like in a p-n junction, this equilibrium is dynamic, not static. However, the difference consists in the fact that in a Schottky barrier current transport is mainly due to majority carriers. In other words, the only way for an electron to overcome the barrier and flow to the other side is by thermionic emission. The current generated by thermionic emitted electrons is given by the Richardson equation [22]:

$$J_{th} = A_0 T^2 exp(-\frac{q\phi_B}{k_B T}), \tag{B.1}$$

where A_0 is the Richardson constant, T is the temperature and ϕ_B is the height of the barrier, given by $\phi_B = \phi_M - \chi_e$. Notice that the height of the barrier is slightly different from the *built-in potential*, defined as $V_{bi} = \phi_M - \phi_S$, making the height

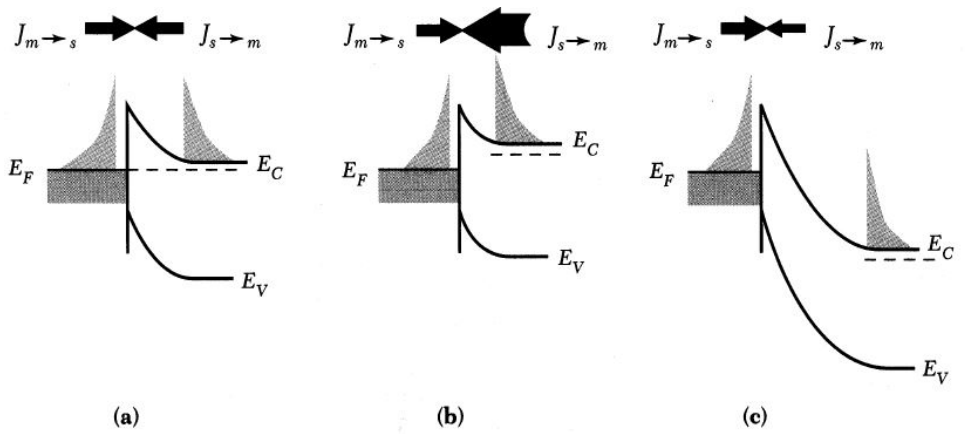


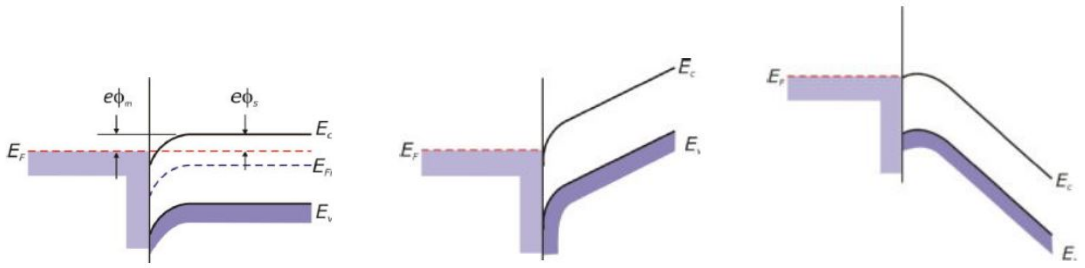
Figure B.1: Schottky contact band diagram under equilibrium (a), with a positive (b) and negative (c) bias applied [22].

of the barrier independent on semiconductor doping. Then, when the junction is brought out of equilibrium by an external bias, if the voltage applied on the metal is positive, the barrier height is reduced and electrons can easily flow from the semiconductor to the metal (Fig.B.1.b). On the contrary, when the bias is negative, the height of the barrier increases and the electrons flow from the semiconductor to the metal is suppressed, whereas very little current flows from the metal to the semiconductor. The result is a exponential I-V relationship, called rectifying because current flow is allowed only in one direction, like in a p-n diode. If the semiconductor is a p-type one, the situation is exactly the same in the situation where $\phi_M < \phi_S$. i.e. in the opposite situation. Schottky barriers are useful for some applications like Schottky diodes, Schottky transistors and MESFETs. However, in most of the cases the desired metal-semiconductor contact is ohmic.

B.2 Ohmic Contacts

An ohmic contact is a contact which shows a linear relationship between the current which flow through it and the voltage drop across it, with a negligible contact resistance. Negligible means that the voltage drop across it has to be very small compared to the total voltage applied to the device contacted through such a contacts (for example, few mV for voltages of few V). This type of contact is essential in semiconductor technology, because a non-linear contact (i.e. Schottky) can distort the output characteristic of a device or, in general, degrade its performance (for example, a lower drain saturation current in a MOSFET). As a consequence, the engineering of an ohmic contact is essential also in the context of this Thesis. The simplest way to achieve an ohmic contact consists in contacting a metal with

a workfunction lower than the semiconductor one, for a n-type semiconductor ($\phi_M < \phi_S$). In this case, in fact, to reach the equilibrium this time electrons move from the metal to the semiconductor. Electrons are accumulated in the semiconductor close to the interface. However, the resulting electric field is weak, being originated not by fixed charges like in the previous case, but by a thin sheet of mobile charges closed to the interface. Also the originated band curvature is too weak to prevent electrons flow through the potential barrier.



(a) **(b)** **(c)**

Figure B.2: Band diagram of a ohmic contact under equilibrium (a), with a positive (b) and negative (c) bias applied [75]

This is exactly the situation reported in Fig.B.2a [75]. Then, if a positive bias is applied, electrons from the semiconductor are attracted into the metal. Their movement is not prevented by any barrier, so they are completely free to flow in that direction (Fig.B.2b. Also when a negative bias is applied, electrons can flow in the opposite direction experiencing a very small barrier. Since no depletion region, i.e. a region with a high resistance, is present, voltage drop is distributed among all the semiconductor (the voltage drop across the metal is always assumed negligible), resulting in a constant electric field and a linear relationship between current and voltage.
However, this is not the only way to realize an ohmic contact. In fact, the same result can be obtained also when $\phi_M > \phi_S$ if the height of the barrier is limited and the depletion region is very narrow. In fact, the width of the potential Schottky barrier extends through the depletion region. If this latter is very narrow, current transport due to thermionic emission is not the prevalent mechanism anymore. Electron passage is not over the barrier, but through the barrier because of quantum-mechanical tunneling effect. The situation is depicted in Fig.B.3 [75]. In conclusion, in a tunneling contact the tunneling probability through the barrier is so high that the current flow is never prevented in any direction because of potential barriers. Carriers flow in one direction or in the other one according to the polarity of the bias voltage. The tunneling probability, in the approximation of a rectangular

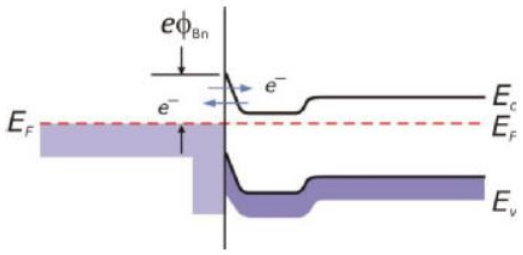


Figure B.3: Band diagram of an ohmic contact obtained by electron tunneling across the potential barrier [75].

barrier, is given by:

$$T = exp(-2k_z z) = exp(-2z\sqrt{\frac{2m_{eff}W}{\hbar^2}}), \tag{B.2}$$

where z is the width of the barrier, k_z the electron wavevector, W the height of the barrier (in J), m_{eff} the effective mass and $\hbar$ the reduced Planck constant. In this case, the height of the barrier corresponds to $W = \phi_B - V$ (being $\phi_B = \phi_M - \chi_e$), whereas the width of the barrier extends over the depletion region:

$$z = x_d = \sqrt{\frac{2\varepsilon_{Si}(\phi_B - V)}{qN_D}}, \tag{B.3}$$

where V is external applied voltage. The tunneling current is proportional to the tunneling probability:

$$I \sim exp(-\frac{C(\phi_B - V)}{\sqrt{N_D}}), \tag{B.4}$$

being $C = 4\sqrt{m_{eff}\varepsilon_{Si}}/\hbar$ a constant. The *specific contact resistance* is defined as [22]:

$$R_C = (\frac{dJ}{dV})^{-1}_{V=0} \sim exp(\frac{C\phi_B}{\sqrt{N_D}}), \tag{B.5}$$

whose measurement unit is $\Omega \cdot cm^2$. So, the contact resistance in Ω is simply given by R_C/A, where A is the area of the contact. In conclusion, to obtain an ohmic contact with a very low contact resistance, semiconductor doping and area of the contact have to be as large as possible, whereas the height of the potential barrier has to be minimized by the choice of the metal.

Bibliography

[1] *5 nm technology*. URL: https://www.tsmc.com/english/dedicatedFoundry/technology/logic/l_5nm (cit. on p. 1).

[2] *IBM Introduces the World's First 2-nm Node Chip*. URL: https://spectrum.ieee.org/nanoclast/semiconductors/nanotechnology/ibm-introduces-the-worlds-first-2nm-node-chip (cit. on p. 1).

[3] *The Nanosheet Transistor Is the Next (and Maybe Last) Step in Moore's Law*. URL: https://spectrum.ieee.org/semiconductors/devices/the-nanosheet-transistor-is-the-next-and-maybe-last-step-in-moores-law (cit. on p. 1).

[4] *The End of More — The Death of Moore's Law*. URL: https://medium.com/@sgblank/the-end-of-more-the-death-of-moores-law-5ddcfd8439dd (cit. on p. 1).

[5] Michael Allan Fogarty. «Spin-Based Quantum Computing with Silicon MOS Quantum Dots». In: (2018) (cit. on pp. 1, 3, 5, 6, 13, 14, 17, 27, 29–31).

[6] *Superposition and Entanglement*. URL: https://www.quantum-inspire.com/kbase/superposition-and-entanglement/ (cit. on pp. 1, 2).

[7] Michael A Nielsen and Isaac Chuang. *Quantum computation and quantum information*. 2002 (cit. on p. 1).

[8] Jelmer Boter. «Scaling Aspects of Silicon Spin Qubits». In: (2020) (cit. on pp. 1, 4–6, 17, 18, 28, 31, 34).

[9] David Morin. *Introduction to quantum Mechanics*. 2008 (cit. on p. 2).

[10] *Wave-particle duality*. URL: https://www.britannica.com/science/wave-particle-duality (cit. on p. 2).

[11] *Copenhagen Interpretation of Quantum Mechanics*. URL: https://seop.illc.uva.nl/entries/qm-copenhagen/ (cit. on p. 2).

[12] *Quantum Entanglement*. URL: https://www.qutube.nl/courses-10/quantum-101/quantum-entanglement (cit. on p. 3).

[13] Lov K Grover. «Quantum mechanics helps in searching for a needle in a haystack». In: *Physical review letters* 79.2 (1997), p. 325 (cit. on p. 3).

[14] Peter W Shor. «Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer». In: *SIAM review* 41.2 (1999), pp. 303–332 (cit. on p. 3).

[15] Richard P Feynman. «Simulating physics with computers». In: *Feynman and computation*. CRC Press, 2018, pp. 133–153 (cit. on p. 3).

[16] David P DiVincenzo. «The physical implementation of quantum computation». In: *Fortschritte der Physik: Progress of Physics* 48.9-11 (2000), pp. 771–783 (cit. on p. 3).

[17] John Chiaverini et al. «Realization of quantum error correction». In: *Nature* 432.7017 (2004), pp. 602–605 (cit. on p. 4).

[18] Daniel Loss and David P DiVincenzo. «Quantum computation with quantum dots». In: *Physical Review A* 57.1 (1998), p. 120 (cit. on pp. 4, 5).

[19] M Field, CG Smith, Michael Pepper, DA Ritchie, JEF Frost, GAC Jones, and DG Hasko. «Measurements of Coulomb blockade with a noninvasive voltage probe». In: *Physical Review Letters* 70.9 (1993), p. 1311 (cit. on p. 5).

[20] Elliot J Connors, JJ Nelson, Haifeng Qiao, Lisa F Edge, and John M Nichol. «Low-frequency charge noise in Si/SiGe quantum dots». In: *Physical Review B* 100.16 (2019), p. 165305 (cit. on pp. 6, 19, 56).

[21] Nathan Holman. *Circuit Quantum Electrodynamics with Si/SiGe Quantum Dots*. The University of Wisconsin-Madison, 2020 (cit. on pp. 6, 19, 26, 31, 32, 34, 56).

[22] SM Sze and Semiconductor Devices. «Physics and technology». In: (1985) (cit. on pp. 6, 10, 11, 34–37, 103, 106, 121, 122, 125, 127, 128, 130).

[23] Neil W Ashcroft, N David Mermin, et al. *Solid state physics*. 1976 (cit. on p. 9).

[24] Thierry Ouisse. *Electron transport in nanostructures and mesoscopic devices: an introduction*. John Wiley & Sons, 2013 (cit. on pp. 9, 12, 13, 20, 21).

[25] Hailing Cheng. «Two-dimensional Electron Transport in Gallium Nitride and Gallium Arsenide-based Heterostructures.» PhD thesis. 2011 (cit. on p. 10).

[26] Tetsu TANAKA, Toshihiro SUGII, and Chenming HU. «New Method of Extracting Inversion Layer Thickness and Charge Profile and Its Impact on Scaled MOSFETs». In: *Extended abstracts of the... Conference on Solid State Devices and Materials*. Vol. 1996. 1996, pp. 10–12 (cit. on p. 12).

[27] Susan J Angus, Andrew J Ferguson, Andrew S Dzurak, and Robert G Clark. «Gate-defined quantum dots in intrinsic silicon». In: *Nano letters* 7.7 (2007), pp. 2051–2055 (cit. on p. 15).

[28] Nard Dumoulin Stuyck et al. «An Integrated Silicon MOS Single-Electron Transistor Charge Sensor for Spin-Based Quantum Information Processing». In: *IEEE Electron Device Letters* 41.8 (2020), pp. 1253–1256 (cit. on p. 16).

[29] WH Lim, CH Yang, FA Zwanenburg, and AS Dzurak. «Spin filling of valley–orbit states in a silicon quantum dot». In: *Nanotechnology* 22.33 (2011), p. 335704 (cit. on pp. 16, 17, 106).

[30] NS Lai, WH Lim, CH Yang, FA Zwanenburg, WA Coish, F Qassemi, A Morello, and AS Dzurak. «Pauli spin blockade in a highly tunable silicon double quantum dot». In: *Scientific reports* 1.1 (2011), pp. 1–6 (cit. on pp. 17, 18).

[31] M Veldhorst et al. «An addressable quantum dot qubit with fault-tolerant control-fidelity». In: *Nature nanotechnology* 9.12 (2014), pp. 981–985 (cit. on p. 19).

[32] J-S Kim, Thomas M Hazard, Andrew A Houck, and Stephen A Lyon. «A low-disorder metal-oxide-silicon double quantum dot». In: *Applied Physics Letters* 114.4 (2019), p. 043501 (cit. on pp. 19, 24, 32–34, 106).

[33] Adrian M. Ionescu. *Lecture notes from "Nanoelectronics" course, EPFL*. Sept. 2020 (cit. on pp. 25, 38).

[34] Dimitrie Culcer, Xuedong Hu, and S Das Sarma. «Dephasing of Si spin qubits due to charge noise». In: *Applied Physics Letters* 95.7 (2009), p. 073102 (cit. on pp. 32, 34).

[35] L Petit et al. «Spin lifetime and charge noise in hot silicon quantum dot qubits». In: *Physical review letters* 121.7 (2018), p. 076801 (cit. on pp. 33, 34).

[36] Blake M Freeman, Joshua S Schoenfield, and HongWen Jiang. «Comparison of low frequency charge noise in identically patterned Si/SiO2 and Si/SiGe quantum dots». In: *Applied Physics Letters* 108.25 (2016), p. 253108 (cit. on p. 34).

[37] Edward H Nicollian, John R Brews, and Edward H Nicollian. *MOS (metal oxide semiconductor) physics and technology*. Vol. 1987. Wiley New York, 1982 (cit. on pp. 36, 37, 103, 106, 109).

[38] C Robert Helms and Edward H Poindexter. «The silicon-silicon dioxide system: Its microstructure and imperfections». In: *Reports on Progress in Physics* 57.8 (1994), p. 791 (cit. on p. 36).

[39] K Konrad Young. «Short-channel effect in fully depleted SOI MOSFETs». In: *IEEE Transactions on Electron Devices* 36.2 (1989), pp. 399–402 (cit. on p. 37).

[40] Philippe Roche, Jean-Luc Autran, Gilles Gasiot, and Daniela Munteanu. «Technology downscaling worsening radiation effects in bulk: SOI to the rescue». In: *2013 IEEE International Electron Devices Meeting*. IEEE. 2013, pp. 31–1 (cit. on p. 37).

[41] *Smart Cut - Soitec*. URL: `https://www.soitec.com/en/products/smart-cut` (cit. on p. 38).

[42] Bich-Yen Nguyen, George Celler, and Carlos Mazuré. «A Review of SOI Technology and its Applications». In: *Journal of Integrated Circuits and Systems* 4.2 (2009), pp. 51–54 (cit. on p. 38).

[43] Benoit Voisin et al. «Few-electron edge-state quantum dots in a silicon nanowire field-effect transistor». In: *Nano letters* 14.4 (2014), pp. 2094–2098 (cit. on p. 39).

[44] Romain Lavieville et al. «Quantum dot made in metal oxide silicon-nanowire field effect transistor working at room temperature.» In: *Nano letters* 15.5 (2015), pp. 2958–2964 (cit. on p. 39).

[45] Fabio Ansaloni, Anasua Chatterjee, Heorhii Bohuslavskyi, Benoit Bertrand, Louis Hutin, Maud Vinet, and Ferdinand Kuemmeth. «Single-electron operations in a foundry-fabricated array of quantum dots». In: *Nature communications* 11.1 (2020), pp. 1–7 (cit. on p. 39).

[46] S De Franceschi et al. «SOI technology for quantum information processing». In: *2016 IEEE International Electron Devices Meeting (IEDM)*. IEEE. 2016, pp. 13–4 (cit. on p. 39).

[47] R Maurand et al. «A CMOS silicon spin qubit». In: *Nature communications* 7.1 (2016), pp. 1–6 (cit. on p. 39).

[48] R. B. Darling. *Oxidation of Silicon, EE-527*. Dec. 2013 (cit. on pp. 42, 106).

[49] Bruce E Deal and AS Grove. «General relationship for the thermal oxidation of silicon». In: *Journal of applied physics* 36.12 (1965), pp. 3770–3778 (cit. on p. 44).

[50] *RCA Cleaning*. URL: `https://www.epfl.ch/research/facilities/cmi/equipment/thin-films/plade-rca/` (cit. on p. 44).

[51] M. Cocuzza. *Lecture notes from "Physics of technological process" course, PoliTO*. Oct. 2019 (cit. on pp. 45, 48, 49, 53).

[52] Ankita Verma, Joshah Jennings, Ryan D Johnson, Marc H Weber, and Kelvin G Lynn. «Fabrication of 3D charged particle trap using through-silicon vias etched by deep reactive ion etching». In: *Journal of Vacuum Science & Technology B, Nanotechnology and Microelectronics: Materials, Processing, Measurement, and Phenomena* 31.3 (2013), p. 032001 (cit. on p. 47).

[53] *Heidelberg Instruments MLA150*. URL: https://www.epfl.ch/research/facilities/cmi/equipment/photolithography/heidelberg-instruments-mla150/ (cit. on p. 48).

[54] *Raith EBPG5000 – Ebeam tool*. URL: https://www.epfl.ch/research/facilities/cmi/equipment/ebeam-lithography/raith-ebpg5000/ (cit. on pp. 48, 59).

[55] John Kangchun Perng. «High aspect-ratio nanoscale etching in silicon using electron beam lithography and deep reactive ion etching (DRIE) technique». PhD thesis. Georgia Institute of Technology, 2006 (cit. on p. 50).

[56] *Electron Beam Source for Electron Beam Deposition*. URL: https://www.jeol.co.jp/en/science/eb.html (cit. on p. 50).

[57] SB Felch, MI Current, and MC Taylor. «Ion implantation for semiconductor devices: The largest use of industrial accelerators». In: *Proceedings of the North American Particle Accelerator Conference*. 2013, p. 740 (cit. on p. 52).

[58] *Disco DAD321*. URL: https://www.epfl.ch/research/facilities/cmi/equipment/packaging-miscellaneous/disco-dad321/ (cit. on p. 53).

[59] *RTP JETFIRST 200*. URL: https://www.epfl.ch/research/facilities/cmi/equipment/thin-films/rtp-jetfirst-200/ (cit. on pp. 53, 54).

[60] N. Garino. *Lecture notes from "Materials and characterization for micro and nanotechnologies" course, PoliTO*. Oct. 2019 (cit. on p. 55).

[61] Matthias Brauns, Sergey V Amitonov, Paul-Christiaan Spruijtenburg, and Floris A Zwanenburg. «Palladium gates for reproducible quantum dots in silicon». In: *Scientific reports* 8.1 (2018), pp. 1–8 (cit. on p. 57).

[62] WIL Lawrie et al. «Quantum dot arrays in silicon and germanium». In: *Applied Physics Letters* 116.8 (2020), p. 080501 (cit. on p. 57).

[63] *Two-layer PMMA e-beam resist system for high-resolution lift-off*. URL: https://www.allresist.com/resist-wiki-two-layer-pmma-e-beam-resist-system-for-high-resolution-lift-off/ (cit. on p. 63).

[64] F. Giorgis. *Lecture notes from "Materials and characterization for micro and nanotechnologies" course, PoliTO*. Oct. 2019 (cit. on p. 66).

[65] *Sopra GES 5E*. URL: https://www.epfl.ch/research/facilities/cmi/equipment/metrology/sopra-ges-5e/ (cit. on p. 70).

[66] Paul A Gagniuc. *Markov chains: from theory to implementation and experimentation*. John Wiley & Sons, 2017 (cit. on p. 80).

[67] Alex Burenkov, Peter Pichler, J. Lorenz, Yohann Spiegel, Julian Duchaine, and Frank Torregrosa. «Simulation of Plasma Immersion Ion Implantation». In: Sept. 2011, pp. 231–234. DOI: 10.1109/SISPAD.2011.6034962 (cit. on p. 82).

[68] *Cascade Probe Station*. URL: https://www.cascademicrotech.com/ (cit. on p. 97).

[69] *Model 4200A-SCS Parameter Analyzer, User's Manual*. URL: https://download.tek.com/manual/4200A-900-01B_User_Manual_Feb2017.pdf (cit. on p. 98).

[70] *C-V Characterization of MOS Capacitors Using the 4200A-SCS Parameter Analyzer*. URL: https://download.tek.com/document/1KW-60645-0_C-V_CharMOSCapacitors_4200A-SCS.pdf (cit. on pp. 98, 99, 125).

[71] Dieter K. Schroder. *Semiconductor Material and Device Characterization*. USA: Wiley-Interscience, 2006. ISBN: 0471749095 (cit. on pp. 99, 121, 122, 124, 125).

[72] Ching-Huang Lu, Gloria MT Wong, Ryan Birringer, Reinhold Dauskardt, Michael D Deal, Bruce M Clemens, and Yoshio Nishi. «Bilayer metal gate electrodes with tunable work function: Mechanism and proposed model». In: *Journal of Applied Physics* 107.6 (2010), p. 063710 (cit. on pp. 102, 103).

[73] Brijesh Kumar, Brajesh Kumar Kaushik, and YS Negi. «Perspectives and challenges for organic thin film transistors: Materials, devices, processes and applications». In: *Journal of Materials Science: Materials in Electronics* 25.1 (2014), pp. 1–30 (cit. on pp. 102, 103).

[74] Diefeng Gu, Sandwip K Dey, and Prashant Majhi. «Effective work function of Pt, Pd, and Re on atomic layer deposited Hf O 2». In: *Applied Physics Letters* 89.8 (2006), p. 082907 (cit. on p. 126).

[75] *Ohmic contact*. URL: https://warwick.ac.uk/fac/sci/physics/current/postgraduate/regs/mpagswarwick/ex5/devices/hetrojunction/ohmic/ (cit. on pp. 129, 130).